British Journal of Social Psychology (2018), 57, 367–385

The British Psychological Society

www.wileyonlinelibrary.com

Bayesian analysis of multimethod ego-depletion studies favours the null hypothesis

Joseph L. Etherton*, Randall Osborne, Katelyn Stephenson, Morgan Grace, Chas Jones and Alessandro S. De Nadai

Texas State University, San Marcos, TX, USA

Ego-depletion refers to the purported decrease in performance on a task requiring self-control after engaging in a previous task involving self-control, with self-control proposed to be a limited resource. Despite many published studies consistent with this hypothesis, recurrent null findings within our laboratory and indications of publication bias have called into question the validity of the depletion effect. This project used three depletion protocols involved three different depleting initial tasks followed by three different self-control tasks as dependent measures (total $n = 840$). For each method, effect sizes were not significantly different from zero When data were aggregated across the three different methods and examined meta-analytically, the pooled effect size was not significantly different from zero (for all priors evaluated, Hedges' $g = 0.10$ with 95% credibility interval of [−0.05, 0.24]) and Bayes factors reflected strong support for the null hypothesis (Bayes factor > 25 for all priors evaluated).

One of the more research-inspiring ideas in recent years is the proposal that self-control is a limited resource, subject to depletion and requiring replenishment (Baumeister & Heatherton, 1996; Heatherton & Baumeister, 1996; Muraven, Tice, & Baumeister, 1998). The proposed reduction in self-control following exertion of self-control on a previous task has been termed 'ego-depletion' (Baumeister & Heatherton, 1996). In this formulation, self-control operates much like a muscle, becoming temporarily weakened after exertion, recovering following rest, and strengthening with practice (Muraven, 2010; Muraven & Baumeister, 2000).

According to this strength model of self-control (Baumeister, Vohs, & Tice, 2007), tasks that involve self-control draw from a shared limited-capacity resource, thereby reducing the quantity of available self-control resources for subsequent tasks. Consequently, exerting self-control in one domain (e.g., resisting sweets while on a diet) would deplete the general self-control resource, and resulting in reduced self-control success in another behaviour that requires self-control resources (e.g., persisting at solving difficult anagrams), even if the second task was otherwise dissimilar to the initial task (e.g., Muraven & Baumeister, 2000; Muraven *et al.*, 1998).

Because self-control is believed to draw from a limited-capacity resource utilized by all self-control activities, any self-control task should be able to serve as the initial depleting task, with any other self-control task as the dependent measure. For example, participants

*Correspondence should be addressed to Joseph L. Etherton, Department of Psychology, Texas State University, 601 University Drive, San Marcos TX 78666, USA (email: je27@txstate.edu).

DOI:10.1111/bjso.12236

asked to refrain from eating nearby snacks (depleting self-control task) consumed significantly more ice cream (dependent measure) in a subsequent taste test (Vohs & Heatherton, 2000). Other depleting self-control tasks include regulating affect while watching emotionally provocative videos (e.g., Baumeister, Bratslavsky, Muraven, & Tice, 1998; Finkel & Campbell, 2001), following complex rules for crossing out letters in a passage of text (e.g., Baumeister *et al.*, 1998; DeWall, Baumeister, Gailliot, & Maner, 2008), performing arithmetic while experiencing auditory interference (e.g., Alberts, Martijn, Greb, Merckelbach, & de Vries, 2007), or suppressing stereotypes about others (e.g., Gailliot, Plant, Butz, & Baumeister, 2007). Dependent measures on which impaired performance has been observed include success at solving math problems (e.g., Johns, Inzlicht, & Schmader, 2008; Tyler & Burns, 2008), performance on a modified Stroop task (e.g., Fennis, Janssen, & Vohs, 2009; Inzlicht, McKay, & Aronson, 2006), and persistence at an unsolvable geometric tracing task (e.g., Baumeister *et al.*, 1998; Fennis *et al.*, 2009); many other tasks have been described in this literature.

Within our own laboratory, however, a number of depletion studies have resulted in null findings. Discussions with other researchers indicated that difficulties obtaining the depletion effect were more common than was suggested by the published literature, raising questions about publication bias.

Subsequently, we planned a series of depletion studies to examine empirically the frequency of null findings, and to estimate the extent of publication bias. Planning for this study occurred during 2011 and 2012, with data collection taking place 2013 to 2016. During planning, the most comprehensive reference for both conceptual and methodological issues was a meta-analysis of depletion studies (Hagger, Wood, Stiff, & Chatzisarantis, 2010). We developed protocols for depleting tasks and outcome measures based on the conceptual underpinnings of the depletion effect as discussed in this meta-analysis, and by reviewing task protocols from the referenced studies.

The Hagger *et al.* (2010) meta-analysis reported an estimated depletion effect size of $d = .62$, which we used to conduct a power analysis to establish sample size ($n = 33$) at 0.80 power. We established $n = 35$ per condition for our studies, which is slightly larger than the average sample size in the published depletion literature of $n = 27$ (Lurquin *et al.*, 2016), because the intent of this series of studies was to examine empirically the frequency of null findings in studies similar in size and procedures to those in the published literature. The use of a similar sample size allows for an empirical examination of how frequently the depletion effect occurs when using methods and sample size similar to those in the pre-2013 published literature.

We developed three different types of depleting tasks and dependent measures similar to tasks that had appeared in at least two published studies referenced in the Hagger *et al.* meta-analysis. The depleting tasks (restricted writing, letter cancellation, and a modified Stroop task) were each paired with different outcome measures (handgrip duration, solving anagrams, and mental arithmetic problems), forming a total of three distinct protocols for inducing and quantifying depletion. We ran each of the three protocols a total of four times, for a total of 12 experimental tests of the depletion effect. This design serves as a more direct approach to estimating the frequency of null findings relative to positive findings and also allows for computation of an effect size estimate based on the pooled results from the 12 separate experiments.

During data collection (2013–2016), several papers critical of the depletion literature were published. The Hagger *et al.* (2010) meta-analysis was criticized on methodological grounds, including insufficient consideration of unpublished null findings (Carter & McCullough, 2013). As one element of their critique, a recently described 'Incredibility

Index' (Schimmack, 2012) was applied to the meta-analysis. This index is derived from the average statistical power of a series of studies to estimate the likelihood that fewer null findings than predicted are present in the group of studies. Assuming a true effect size of $d = .62$ as reported in the Hagger *et al.* (2010) meta-analysis, Carter and McCullough (2013) calculated an average power estimate of 0.55 for these studies. Of the 198 separate experiments reviewed, only 47 reported non-significant findings. Calculations based on the Incredibility Index led the authors to conclude that 'the probability of drawing a set of 198 experiments in which only 47 or fewer were non-significant is roughly 3.7 in one billion' (Carter & McCullough, 2013; p. 683). They concluded that small-study effects likely biased the results of the meta-analysis; that many null or negative studies likely existed but remained unpublished due to publication bias; and that the meta-analysis therefore significantly overestimated the depletion effect size. In their correction for the presumed missing null findings, the confidence interval for the depletion effect size included zero, consistent with the non-existence of the ego-depletion effect. They advised that 'as things stand, we believe that the highest priority for research on the depletion effect should not be arriving at a better theoretical account, but rather determining with greater certainty whether an effect to be explained exists at all' (Carter & McCullough, 2013, p. 684).

A similar criticism was applied to a series of studies involving the role of blood glucose in the depletion effect task (Gailliot, Baumeister, *et al.*, 2007). Performing a self-control task reportedly led to diminished blood glucose, which was associated with poorer subsequent performance on a separate self-control task. Further, ingestion of glucose resulted in normal performance on a subsequent self-control task, whereas those who did not receive glucose performed significantly more poorly on the second self-control task, a finding that was replicated a number of times (e.g., DeWall *et al.*, 2008; Gailliot, Peruche, Plant, & Baumeister, 2009; Masicampo & Baumeister, 2008). However, in evaluating the nine studies reported by Gailliot *et al.* (2007), Lange and Eggert (2014) calculated a *total* power across the nine studies (i.e., the likelihood of obtaining only significant effects across all studies) of less than 1%. They concluded that the reported depletion effect sizes were inflated, presumably reflecting publication bias. Notably, a glucose replication study with greater statistical power found no depletion effect (Dvorak & Simons, 2009).

Another group of researchers developed a Registered Replication program (Hagger *et al.*, 2015), in which a single depletion task protocol was developed and administered by a total of 23 independent laboratories, with the methods, hypotheses, and planned analyses publicly specified in advance of data collection. Of the 23 replications attempted, 20 resulted in effect size confidence intervals that included the value zero. Two replications were statistically significant in the direction opposite that predicted by ego-depletion theory. Only one replication was in the direction of the hypothesized depletion effect (Hagger *et al.*, 2015). In a separate preregistered study utilizing a different methodology (a video-viewing attention control task for depletion), no depletion effect on subsequent operation-span task performance was observed (Lurquin *et al.*, 2016). Bayesian analysis of the results favoured the null hypothesis.

Carter and colleagues conducted their own series of focused meta-analytic tests that avoided methodological limitations of previous work (Hagger *et al.*, 2010), and concluded that there is little evidence to support the existence of the depletion effect (Carter, Kofler, Forster, & McCullough, 2015). Although they observed a statistically significant depletion effect under some analytic conditions, they concluded that the overall pattern of results suggested that the depletion effect was not robust and was not significantly different from zero (Carter *et al.*, 2015). However, this meta-analysis was itself criticized on methodological grounds by Cunningham and Baumeister (2016), who observed that replication failures may be attributable in part to methodologically weak studies conducted by inexperienced researchers, as evidenced by a high proportion of graduate student authorship among the unpublished depletion studies included in the Carter *et al.* (2015) meta-analysis. In addition, the authors criticized a failure to address research quality, and failure to assess whether tasks were adequately operationalized. Further, they argue that Carter and colleagues provided inadequate justification for excluding the majority of the 620 studies that they considered, including only 116 articles in their meta-analyses. They argue the Carter *et al.* meta-analyses are inadequate to draw conclusions about the depletion literature.

Thus, since commencing our project, a number of researchers have published methodological and empirical papers providing substantial evidence of publication bias in the depletion literature, and raising doubts about the legitimacy of the depletion effect. At the same time, some of these criticisms have in turn been criticized on the basis of their own methodological weaknesses. The current study contributes to this literature by reporting the outcome of a series of 12 consecutive studies conducted with sample sizes and procedures consistent with the published depletion literature as of 2013. This project is not intended as a replication of any particular previous studies or of specific methodologies, nor is it intended to characterize potential boundary characteristics of depletion effects. Rather, this project was motivated by the hypothesis that researchers who attempted novel depletion studies but obtained null results often either did not submit or were unsuccessful in publishing their results. As such, this project is intended as a report of the outcome of multiple studies using methods conceptually consistent with the published depletion literature, and with a comparable sample size.

Method

All studies were conducted at a large university in the south-western United States. All participants were recruited either from the Psychology Department subject pool for fulfilment of course requirements or offered extra credit as incentive. Inclusion criteria included normal or corrected-to-normal vision, and intact hearing sufficient to comprehend oral instructions. This research was approved by the Institutional Review Board, and participants completed informed consent forms prior to participation. All measures, manipulations, and exclusions in these studies are reported.

Although we conducted an *a priori* power analysis to determine adequate sample size based on the moderate effect size (d = .62) reported by Hagger *et al.* (2010), which indicated that $n = 33$ per condition was sufficient to detect an effect at $p = .05$ (one-tailed), with 0.80 power, this sample size would be considered inadequate in the light of much smaller effect size estimates recently published. However, our final sample size of $n = 35$ per condition, slightly larger than the average sample size of $n = 27$ per condition (Lurquin *et al.*, 2016) in depletion studies, is entirely consistent with and adequate for the underlying purpose of the study, which is to observe the frequency of null findings in multiple iterations of depletion studies similar to those published prior to 2013. Given that the great majority of published studies prior to 2013 reported significant depletion effects while utilizing sample sizes averaging $n = 27$ per condition, our series of studies serves as a direct examination of the frequency with which significant versus null findings obtain when all results are reported. This sample size was established prior to initiating data collection, and data collection was neither extended nor abbreviated on the basis of the results of data analyses.

Task selection

Because the conceptual basis for the depletion effect is that engaging in self-control draws from a limited-capacity resource that underlies, and is depleted by, all self-control tasks, we selected tasks from a variety of domains. A lengthy appendix (Hagger *et al.*, 2010) listing depleting tasks and dependent measures that have appeared in published studies was consulted. We sought tasks that involved different domains, such as persistence at a motor task (e.g., handgrip duration), as well as cognitive tasks (e.g., Stroop). Although we were not attempting to replicate any one specific technique or prior study, we selected types of tasks that were consistent with the principles of depletion research, in that they involved the exertion of some degree of self-control.

Although there is some disagreement among researchers regarding which tasks involve primarily self-control, one of the variables most closely associated with depletion tasks, both conceptually and empirically, is perceived difficulty, and method sections typically describe features of the depleting task that make it more difficult than the non-depleting control task. Indeed, the depletion effect has been observed in tasks that were rated as *difficult* but had not conventionally been viewed as self-control tasks, such as solving multiplication problems (Hagger *et al.*, 2010). In addition, the meta-analytic effect size of differences between depleted and control groups was greater for the variable 'perceived difficulty' (d = .94) than for other variables associated with depletion, including 'effort' (d = .64), 'fatigue' (d = .44), and 'blood glucose' (d = −.87) (Hagger *et al.*, 2010). As such, we identified the construct 'difficulty' as a reasonable index for determining whether a task would be depleting, and collected difficulty ratings for all depleting tasks and control tasks. For each method in the current series of studies, difficulty ratings for the depleting tasks were significant higher than for the control tasks, with large effect sizes.

General procedure

For each study, an Initial task was administered with instructions that either required self-control (Depletion) or did not (Control). For each study, a different self-control task, referred to here as the Outcome task, was then administered to measure the degree of depletion caused by the first task. Following completion of both tasks, participants were asked to numerically rate the difficulty of the initial depleting task. In all cases, a single experimenter administered tasks to each participant individually in a single session of approximately 20 min. On arrival, participants reviewed and signed a consent form, and were then assigned via restricted randomization (sample size set at 35 per condition) to either a Depletion or Control condition. Once data had been gathered from 35 participants in each of the Depletion and Control conditions (total n = 70), data were demarcated from the next 70 participants (e.g., Study 1A from Study 1B), and so on, until a total of 4 studies (1A through 1D) had been completed. After all four studies for the first protocol (crossing-out letters task and anagrams) had been completed (i.e., 280 total participants across four studies [1A–1D] of this methodology), the next protocol (colour naming and grip squeeze duration) was started, following the same procedure of running 70 participants per study for four (2A–2D) separate studies. The third protocol (writing exercise and math problems) was then administered following the same procedures, also with four studies (3A–3D).

STUDIES 1A–1D

Participants

A total of 280 undergraduate volunteers were recruited via class emails as well as online posting on a Psychology Department research requirement website. Missing data for one participant each for the Control and Depletion conditions resulted in a final total sample size of 278 (192 female, 86 male; mean age = 19.6). The two conditions did not differ in mean age or in gender distribution. Restricted random assignment (setting $n = 35$ per condition) was used to place the first 70 participants in either the Depletion condition or the Control condition. Completion of data collection for the first 35 participants per condition marked the end of Study 1A, after which the next 70 participants were randomly assigned as above, marking Study 1B, and so on, until four separate studies had been administered.

Materials and procedure

Depleting task

For Studies 1A–1D, we developed a crossing-out letters task similar to the one that has been used in previous studies (Baumeister *et al.*, 1998). Participants in both conditions were provided with a lengthy passage of text. Those in the Control condition simply crossed out with a pencil all occurrences of the letter 'e' in a passage of text, whereas those in the Depletion condition were given the following instructions: '*cross off the letter "e" every time it appears with the following exceptions: 1. Do not cross out the "e" if it is adjacent to another vowel (e.g., friend); 2. Do not cross out the "e" if it is one letter away from another vowel (e.g., vowel); 3. Do not cross out the "e" if the word has 6 letters (e.g., "there"); 4. Do not cross out the "e" if it is the third to last letter (e.g., customers); 5. Do not cross out the "e" if there are double letters in the word (e.g., "hello")*'. Participants were asked to continue the writing task for six minutes, after which the experimenter instructed participants to discontinue. We note that subsequent papers using this task have described a habit-reversal manipulation, such that those in the Depletion condition would first cross out the letter 'e' whenever encountered, after which they would then be asked to follow the complex rules described above for crossing out the letter 'e' (e.g., DeWall *et al.*, 2008). However, no such instructions appear in the original description of this task, nor is there any introduction of the role of habit reversal in contributing to depletion effects in the original description of this manipulation. Rather, the original description of the task notes that self-regulatory difficulty was increased by requiring participants to consult multiple rules, so that 'the task was made quite difficult' (Baumeister *et al.*, 1998; p. 1260); habit reversal was not invoked in the original formulation of this task (see Baumeister *et al.*, 1998; pp. 1259–1260). As such, the current task is methodologically consistent with the original task and conceptually consistent with resource depletion as a task requiring increased self-regulation. Other researchers using a crossing-out letters task without a habit-reversal manipulation reported greater effort ratings for the complex rules task and observed significant depletion effects (Wan & Sternthal, 2008). For the current study, difficulty ratings served as a manipulation check and confirmed that the depletion task was more difficult than the control task.

Outcome task

An anagram-solving task similar to previously described tasks (Baumeister *et al.*, 1998; Gailliot *et al.*, 2007; Park, Glaser, & Knowles, 2008) was administered immediately following the crossing-out letters task. This task has been used in a number of self-regulation studies, requiring self-regulatory engagement by creating and then overriding various letter combinations until a solution is reached (e.g., Gailliot *et al.*, 2007). Participants were given a sheet on which 30 anagrams were listed. Participants were asked to complete as many anagrams as they could within five minutes. After the anagram task, participants were asked to refer to a Likert scale to rate the difficulty of the initial crossing-out letters task (1 = not at all difficult; 7 = very difficult). The number of correctly completed anagrams was recorded for each participant.

Results

To confirm that the depletion task was experienced as more difficult, and therefore more demanding of self-control than the control task, difficulty ratings were compared by condition. For all of Studies 1A through 1D, those in the Depletion condition rated their letter-cancellation task as significantly more difficult than those in the Control condition (see Table 1).

Separate independent t-tests and effect sizes (Cohen's *d*) for each of Studies 1A through 1D were conducted, with the number of anagrams correctly solved serving as the dependent variable. Table 2 presents the results of these analyses as well as descriptive statistics by condition. No statistically significant differences in anagrams solved were observed for any of Studies 1A through 1D. Because the methods used and participant source were the same for all of Studies 1A–1D, we pooled data across all four studies ($n = 278$). An independent-samples t-test for the pooled data (Control condition $M = 6.70$, $SD = 2.05$; Depletion condition $M = 6.30$, $SD = 1.90$) revealed no depletion effect, $t(276) = 1.67$, $p = .097$, Cohen's $d = .20$; CI_{95} [−0.04, 0.43].

A Bayesian analysis was performed to test the relative likelihood of observing our results given the null hypothesis (i.e., the absence of any depletion effect) versus the likelihood of observing our results given the alternative hypothesis (the presence of a nonzero effect). A Bayes factor value is a ratio reflecting the likelihood of one hypothesis against a different hypothesis, given the observed data. A Bayes factor of 1 indicates that

Table 1. Crossing-out letters task difficulty ratings by condition for Studies 1A–1D.

	Control		Depletion				
Study	*M*	*SD*	*M*	*SD*	*t* (68)	*p*	ES (95% CI)
1A (*n* = 69)	2.80	1.28	4.44[a]	1.13	5.64[b]	.000	1.36 (0.82, 1.86)
1B (*n* = 70)	2.94	1.39	4.66	1.16	5.59	.000	1.34 (0.81, 1.85)
1C (*n* = 70)	2.77	1.24	4.46	1.12	5.97	.000	1.43 (0.89, 1.94)
1D (*n* = 69)	3.09[a]	1.53	4.90	1.31	5.43[b]	.000	1.27 (0.74, 1.77)

Notes. ES, Effect size (Cohen's *d*, 95% confidence interval).
n = 35 per condition per study, except where noted.
Difficulty rating scale: 0 = 'not at all difficult' to 7 = 'very difficult'.
[a]*n* = 34 due to dropped one participant each.
[b]For 1A and 1D, degrees of freedom = 67 due to dropped participants.

Table 2. Anagrams solved by condition for Studies 1A–1D.

	Control		Depletion				
Study	*M*	*SD*	*M*	*SD*	*t* (68)	*p*	ES (95% CI)
1A (n = 69)	7.14	1.80	6.56[a]	1.89	1.31[b]	.194	0.27 (−0.21, 0.74)
1B (n = 70)	7.03	2.05	6.46	1.77	1.25	.216	0.26 (−0.22, 0.72)
1C (n = 70)	5.80	1.95	6.06	2.10	0.53	.597	−0.15 (−0.62, 0.32)
1D (n = 69)	6.82[a]	2.20	6.14	1.87	1.39[b]	.169	0.34 (−0.14, 0.81)
Pooled (n = 278)	6.70	2.05	6.30	1.90	1.67[c]	.097	0.20 (−0.04, 0.43)

Notes. SD, standard deviation; ES, effect size (Cohen's *d*, 95% confidence interval).
n = 35 per condition per study, except where noted.
[a]For 1A and 1D, n = 34 due to one dropped participant each.
[b]For 1A and 1D, degrees of freedom = 67 due to one dropped participant each.
[c]Degrees of freedom for pooled data = 276.

the observed data do not favour one hypothesis over the other, Bayes factor values from 1 to 3 are interpreted as 'weak' support for the alternative hypothesis over a null hypothesis, and values of 3 to 10 are interpreted as 'strong' support for the alternative hypothesis (Jarosz & Wiley, 2014), with increasing Bayes factor values indicating increasingly stronger support.

The JZS prior, which combines a Cauchy distribution as the effect size prior, and the Jeffreys prior on variance (Jeffreys, 1961), has been recommended as a broadly non-informative prior that minimizes the weighting of implausible and unreasonable values (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Pooled data across the four studies were submitted to the online Bayes factor calculator at http://pcl.missouri.edu/bayesfactor (Rouder *et al.*, 2009). The results (scaled JZS Bayes factor = 2.02, default prior scale parameter r = .707) indicated that the observed data were more likely under the null hypothesis rather than the alternative hypothesis by a ratio of 2.02 to 1.

STUDIES 2A–2D

Participants

A total of 280 undergraduate volunteers (201 female, 79 male; mean age = 20.0) were recruited and assigned to conditions with procedures identical to those in Studies 1A–1D, resulting in four separate studies of 70 participants each. The two conditions did not differ in either mean age or gender distribution. For Study 2, the Initial task, modified Stroop colour-naming task similar to those described in previous studies (Fennis *et al.*, 2009; Gailliot *et al.*, 2007) was followed by a handgrip squeeze duration Outcome task. Although some researchers have incorporated the use of a baseline handgrip duration prior to administering the depletion task as a means of controlling for individual differences in grip strength (e.g., Inzlicht *et al.*, 2006; Muraven *et al.*, 1998), we did not do so out of concern that the act of measuring baseline handgrip duration in the same experimental session would in itself require self-control and so would cause depletion in both conditions. That is, because maximum handgrip duration is itself a depletion task, both conditions would subsequently have been depleted immediately prior to the study manipulation of interest, thus establishing two depleted groups instead of one control group and one depletion group. Furthermore, duration of handgrip squeeze has been found to be uncorrelated with either maximum grip strength or overall bodily strength, and instead reflects primarily self-regulation (Muraven *et al.*, 1998), obviating the conceptual basis for administering a baseline measure of handgrip duration. As such, we modified this procedure by eliminating the potentially confounding baseline handgrip duration task.

Materials and procedure

Depleting task

Participants completed an initial Stroop-like colour-naming task in which they were given a stapled set of five pages on which colour names were listed. For those in the Control condition, the colour names were printed in black letters, and they were instructed to simply name the colour word as written, a low-effort task involving only well-practiced reading abilities. Those in the Depletion condition were given modified Stroop instructions similar to those described by Wallace & Baumeister (2002). Standard Stroop task instructions are to name the colour of ink in which a word is printed, which requires self-control (inhibition of a prepotent response), when the written word conflicts with the ink colour in which it is written (e.g., '*Green*' written in blue font). The modified Stroop instructions in the current study were to name the colour of the ink in which words were written, except that for words printed in the ink colour red, they were to name the word that was printed in red. This modification increases task complexity by requiring switching between task instructions for different stimulus types, requiring additional self-control resources and presumably enhancing depletion effects. The experimenter referred to a scoring key to record response accuracy during the task. The colour-naming task was terminated by the experimenter at 3 min 40 s, similar to previously published methods (Bray, Ginis, Hicks, & Woodgate, 2008). Difficulty ratings served as a manipulation check to ensure that the Depleting task was demanding of self-control resources (see Table 3—difficulty rating).

Outcome task

Following the colour-naming task, the handgrip duration task outcome measure was administered. Participants were given a commercially available handgrip exerciser which had been purchased from a sporting goods store. Following experimenter directions, they

Table 3. Colour-naming task difficulty ratings by condition for Studies 2A–2D.

	Control		Depletion				
Study	*M*	*SD*	*M*	*SD*	*t* (68)	*p*	ES (95% CI)
2A (*n* = 70)	2.86	1.35	5.09	1.22	7.23*	.000	1.73 (1.17, 2.26)
2B (*n* = 70)	3.17	1.36	5.14	0.85	7.28*	.000	1.74 (1.17, 2.27)
2C (*n* = 70)	3.46	1.30	5.17	1.22	5.65*	.000	1.36 (0.82, 1.86)
2D (*n* = 70)	3.11	1.23	4.97	1.01	6.89*	.000	1.65 (1.09, 2.18)

Notes. *SD*, standard deviation; ES, effect size (Cohen's *d*, 95% confidence interval).
n = 35 per condition per study, except where noted.
Difficulty rating scale: 0 = 'not at all difficult' to 7 = 'very difficult'.
*$p < .05$

held the handgrip vertically and squeezed it closed, while the experimenter placed a card between the two handles of the handgrip device, where the card was held against gravity as long as the handgrip was held closed. Participants were asked to squeeze the handgrip closed against resistance for as long as they could. The experimenter recorded in seconds the time from the start of the task until the card dropped from the closed handgrip (i.e., when the participant was no longer squeezing the handgrip closed). Participants were not given specific duration goals other than to maintain the handgrip in a closed position for as long as they could. After the handgrip task, participants were asked to rate the difficulty of the initial colour-naming task.

Results

For all of Studies 2A through 2D, difficulty ratings for the modified Stroop task in the Depletion condition were significantly higher than difficulty ratings for the colour word-naming task in the Control condition, with large effect sizes (see Table 3).

For each of Studies 2A through 2D, differences in grip duration performance were evaluated via separate independent-samples t-tests. Table 4 presents the results of these analyses. No statistically significant differences in duration were observed for any of the four studies. When data were pooled across the four studies, an independent-samples t-test (Control condition $M = 52.46$, $SD = 42.89$; Depletion condition $M = 55.02$, $SD = 47.58$) revealed no significant difference, t (278) = 0.47, $p = .637$, Cohen's $d = -.06$, CI_{95} [−0.29, 0.18].

A Bayesian analysis was performed as for pooled data from Study 1, the results of which indicated that the data strongly favoured the null hypothesis (scaled JZS Bayes factor = 6.85, $r = .707$), such that the observed data were more likely under the null hypothesis rather than the alternative hypothesis by a ratio of 6.85 to 1.

STUDIES 3A–3D

Participants

A total of 280 undergraduate volunteers were recruited and assigned to conditions identically as in Studies 1A–1D and 2A–2D above, resulting in four separate studies of 70 participants each ($n = 35$ each for the Control and Depletion conditions). Incomplete

Table 4. Grip squeeze duration in seconds by condition for Studies 2A–2D.

	Control		Depletion				
Study	*M*	*SD*	*M*	*SD*	*t* (68)	*p*	ES (95% CI)
2A (*n* = 70)	52.46	46.78	49.64	45.99	0.25	.800	0.06 (−0.41, 0.53)
2B (*n* = 70)	46.84	31.70	60.74	49.96	1.39	.169	−0.33 (−0.80, 0.14)
2C (*n* = 70)	62.64	47.63	65.67	54.76	.25	.806	−0.06 (−0.53, 0.41)
2D (*n* = 70)	47.89	43.56	44.02	36.65	.40	.689	0.10 (−0.37, 0.56)
Pooled (*n* = 280)	52.46	42.89	55.02	47.58	.47[a]	.637	−0.06 (−0.29, 0.18)

Notes. *SD*, standard deviation; ES, effect size (Cohen's *d*, 95% confidence interval).
n = 35 per condition per study, except where noted.
[a]Degrees of freedom for pooled data = 278.

data resulted in four participants being removed from the Control condition, resulting in a final sample size of 276 (193 female, 80 male, three unspecified; mean age = 19.71). The two conditions did not differ by mean age or gender distribution.

Materials and procedure

Depletion task

Participants completed an initial writing task similar to that described by Schmeichel (2007) in which they were asked to spend 6 min writing a description of a recent trip. For those in the Control condition, there were no constraints on their writing task. Those in the Depletion condition were given the following written instructions: '*Do not use the letters **A** or **N** anywhere in your story! If you find yourself writing a word that includes the letters **A** or **N**, please stop writing that word and find an alternate way to express your thoughts*'. This manipulation requires self-control in the form of inhibiting overlearned words and phrases and effortfully searching for words that are appropriate for the passage but do not include the excluded letters. Difficulty ratings served as a manipulation check to ensure that the Depleting task was demanding of self-control resources (see Table 5—difficulty rating)

Outcome task

On completion of the writing task, participants were given four sheets of paper stapled together with a total of 254 single-digit multiplication problems, with a blank space after each for participants to write their solutions. Solving mathematical problems has been used as an outcome measure in several previous depletion studies (Johns *et al.*, 2008; Tyler & Burns, 2008; Vohs, Baumeister, & Ciarocco, 2005; Wright *et al.*, 2007; Mead, Baumeister, Gino, Schweitzer, & Ariely, 2009). In the Hagger *et al.* (2010) meta-analysis, solving single-digit multiplication problems was regarded as lower in complexity compared with other mathematical tasks, and so was classified as a 'simple self-control task' (Hagger *et al.*, 2010), with self-control likely required in the form of maintaining task

Table 5. Writing task difficulty ratings by condition for Studies 3A–3D.

	Control		Depletion				
Study	*M*	*SD*	*M*	*SD*	*t* (68)	*p*	ES (95% CI)
3A (*n* = 70)	1.91	1.67	6.17	.75	13.77*	.000	3.29 (2.54, 3.97)
3B (*n* = 70)	2.09	2.02	6.20	.87	11.07*	.000	2.64 (1.98, 3.25)
3C (*n* = 66)	1.45[a]	1.71	5.94	1.11	12.80*[b]	.000	3.16 (2.40, 3.84)
3D (*n* = 70)	2.00	1.80	6.06	1.29	10.41*[b]	.000	2.59 (1.93, 3.20)
Pooled (*n* = 276)	1.87	1.80	6.09	1.01	23.82*[c]	.000	2.90 (2.56, 3.23)

Notes. *SD*, standard deviation; ES, effect size (Cohen's *d*, 95% confidence interval).
n = 35 per condition per study, except where noted.
Difficulty rating scale: 0 = 'not at all difficult' to 7 = 'very difficult'.
[a]*n* = 31 due to incomplete data.
[b]For 3C and 3D, degrees of freedom = 64 and 63, respectively, due to missing difficulty rating data.
[c]Pooled degrees of freedom = 269 due to missing difficulty rating data.
*$p < .05$

pace and persisting at the moderately difficult task rather than ceasing. However, despite being identified as lower in complexity, meta-analyses indicate that depletion effect sizes were not significantly different for complex versus simple dependent measures (Hagger *et al.*, 2010). In previous studies, this task was rated as moderately difficult ($M = 4.91$ on a 0–10 difficulty rating scale), and it was shown to be impaired by previously performing a depleting task (Wright, Stewart, & Barnett, 2008). We set a 3-minute time limit for completing as many mathematical problems as possible, similar to previously published methods (Wright *et al.*, 2007, 2008).

Experimenters referred to an answer key to record the accuracy of responses, and the number of problems accurately solved in 3 min served as the dependent measure. On completing both tasks, participants were asked to rate the difficulty of the initial writing task.

Results

Task difficulty ratings were compared by condition. The restricted writing task in the Depletion condition was rated as significantly more difficult than the standard writing task in the Control condition (see Table 5).

For each of Studies 3A through 3D, independent-samples t-tests were conducted to test for differences by condition in solving multiplication problems. Table 6 presents the results of these analyses, including effect sizes and descriptive statistics for number of multiplication problems accurately completed. For Study 3A, the Control group solved significantly more multiplication problems ($M = 92.57$, $SD = 30.03$) than the Depletion group ($M = 76.43, SD = 32.19$), $t(68) = 2.17, p = .034$. However, none of the remaining three studies (3B–3D) revealed any differences by condition in multiplication performance. When data were pooled across all four studies, the overall means for the Control condition ($M = 84.2$, $SD = 33.1$) and the Depletion condition ($M = 79.3$, $SD = 32.6$) were not significantly different, $t(274) = 1.25, p = 0.21, d = .15$, CI_{95} [−0.09, 0.39]. The Bayes factor calculated for the pooled data indicated that the data were in support of the null hypothesis (scaled JZS Bayes factor = 3.60, scale $r = .707$).

Means and standard deviations for each of the 12 studies were evaluated by Bayesian meta-analysis using the package bayesmeta (Roever & Friede, 2017) in R 3.4.2 (R Core

Table 6. Arithmetic problems solved by condition for Studies 3A–3D.

	Control		Depletion				
Study	*M*	*SD*	*M*	*SD*	*t* (68)	*p*	ES (95% CI)
3A (*n* = 70)	92.57	30.03	76.43	32.19	2.17*	.034	0.52 (0.04, 0.99)
3B (*n* = 70)	88.74	35.71	81.00	31.38	.96	.339	0.23 (−0.24, 0.70)
3C (*n* = 66)	76.65[a]	28.40	85.29	36.20	1.07[b]	.289	−0.29 (−1.23, 0.66)
3D (*n* = 70)	77.94	35.54	74.37	30.54	.45	.654	0.11 (−0.36, 0.58)
Pooled (*n* = 276)	84.19	33.05	79.27	32.58	1.25[c]	.214	0.15 (−0.09, 0.39)

Notes. *M*, mean; *SD*, standard deviation; ES, effect size (Cohen's *d*, 95% confidence interval).
n = 35 per condition per study, except where noted.
[a]*n* = 31.
[b]Degrees of freedom = 64.
[c]Degrees of freedom for pooled data = 274.
*$p < .05$

Team, 2014). Hedges' *g* was used as the effect size for all meta-analytic estimates. Given prior research, models with three priors on the mean difference were evaluated: $\mu = 0.26$, $\mu = 0$, and an improper uniform (i.e., uninformative) prior on μ. For all meta-analytic models, a proper uniform prior corresponding to I^2 was used as the prior for τ, and posterior medians of μ were used to define the aggregated mean difference. For prior $\mu = 0.26$, the aggregated mean difference was 0.10 (95% credibility interval: [−0.05, 0.24]), and the Bayes factor for supporting the hypothesis of $\mu = 0$ was 25.94. For prior $\mu = 0$, the aggregated mean difference was again 0.10 (95% credibility interval: [−0.05, 0.24]), and the Bayes factor for supporting the hypothesis of $\mu = 0$ was 25.93. For the model with a uniform prior on μ, the mean difference was again 0.10 (95% credibility interval: [−0.05, 0.24]). Bayes factors were not available for the uniform μ prior model due to the improper priors used in the model. All reported Bayes factors reflect 'strong' support for the null hypothesis according to criteria specified by Kass and Raftery (1995). A forest plot of meta-analytic results based on prior $\mu = 0.26$ is shown in Figure 1.

Discussion

A large body of research has been published on the strength model of self-control and the associated depletion effect. Although the majority of published depletion studies prior to 2013 have reported significant findings, an increasing number of replication failures have been reported (e.g., Hagger *et al.*, 2015; Lurquin *et al.*, 2016; Murtagh & Todd, 2004), and concerns have been raised that the extant literature is likely distorted by publication bias (e.g., Carter & McCullough, 2013, 2014).

In the current study, we have attempted to investigate empirically the extent to which publication bias may have distorted an accurate understanding of hypothesized depletion effects. We did so by conducting multiple studies using methods consistent with the principles of the limited-capacity model of self-control, and similar to methods described in the published literature. We used a sample size ($n = 35$) similar to that of depletion studies conducted prior to 2013, for which the mean sample size per condition was approximately $n = 27$. As such, our series of studies serves as an empirical examination of how frequently null findings actually obtain when using methods and sample sizes similar to those in the published depletion literature. Our results suggest that more depletion studies with null findings likely occurred than would be suggested by the published literature.

However, the relevance of our studies rests on the adequacy of our tasks. To adequately test depletion theory, our initial depletion tasks would need to engage self-control resources sufficiently to induce a depleted state, and the dependent measures would need to demand self-control resources enough to be affected by depletion. Achieving certainty about the adequacy of depleting tasks and dependent measures is difficult given the lack of consensus about precisely which tasks involve a high versus minimal degree of self-control. The depletion literature includes dozens of different ostensible self-control tasks, but a relative paucity of standardization of the parameters (e.g., task duration and intensity). The field would likely benefit from such efforts.

We argue that each of our initial depleting tasks (crossing-out letters, modified Stroop, and restricted writing) did engage self-control sufficiently. Our crossing-out letters protocol was essentially identical to the first published description of this task (Baumeister *et al.*, 1998). Although subsequent researchers have utilized a habit-reversal paradigm (both conditions first cross out all 'es' in text, after which the Depletion condition follows

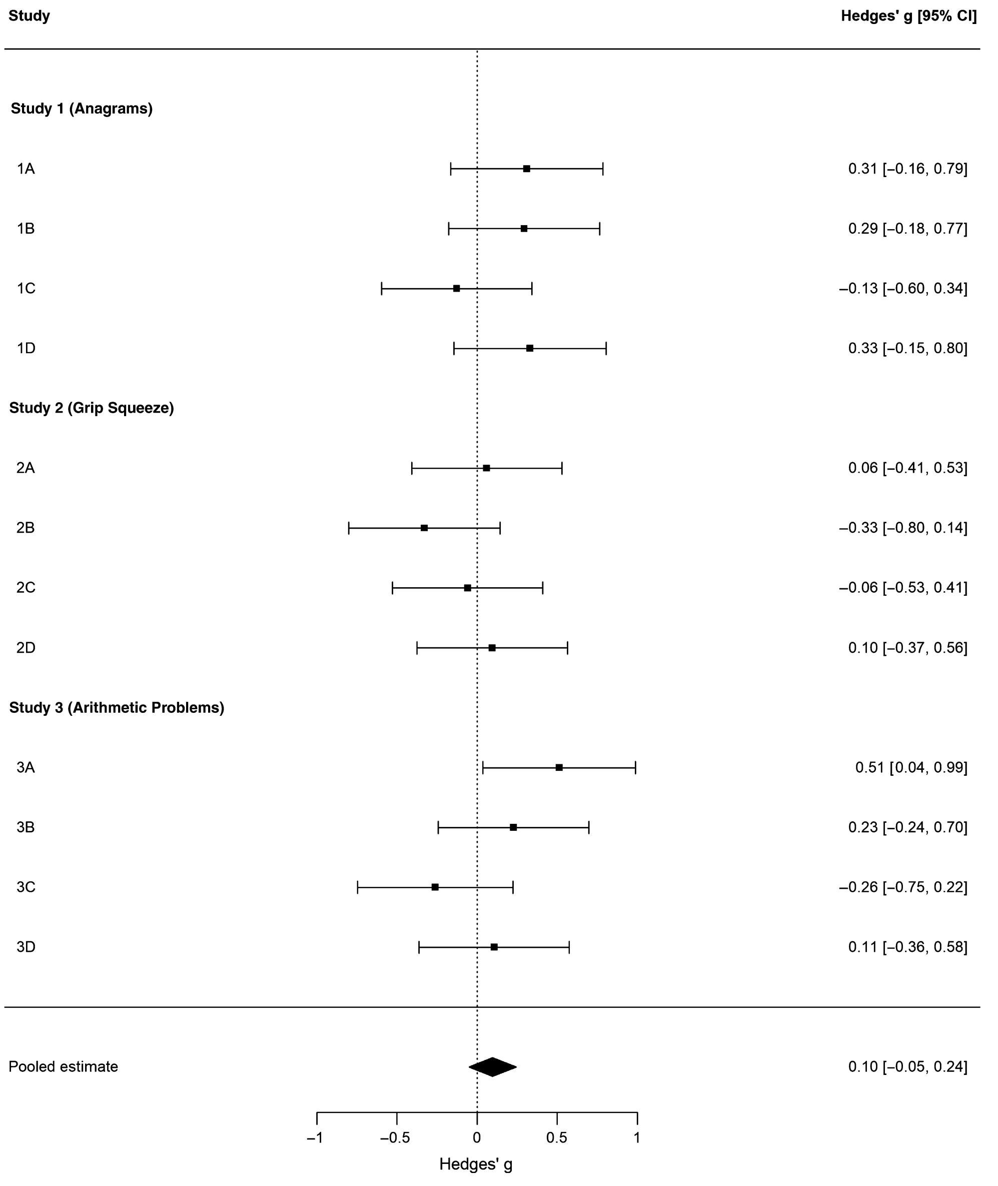

Figure 1. Forest plot of effect sizes (in the metric of Hedges' g) along with 95% confidence intervals for individual studies and 95% credibility interval for the pooled estimate (based on prior $\mu = 0.26$).

complex rules), examination of the original method indicates no such procedures or rationale, noting instead that 'for… the depletion condition, the task was made quite difficult, requiring them to consult multiple rules and monitor their decisions carefully' (Baumeister *et al.*, 1998, p. 1260).

The other depleting tasks, modified Stroop and restricted writing, have been relatively widely accepted as engaging self-control. For the crossing-out letters task, the mean effect size (pooled across Studies 1A–1D) for differences in difficulty rating by condition was Cohen's $d = 1.35$, which is comparable to the mean difficulty rating effect size for the modified Stroop task, Cohen's $d = 1.62$. As noted earlier, task difficulty has been closely aligned with the depletion effect, with depleted versus non-depleted effect sizes larger for difficulty ratings than for any other measured variable (Hagger *et al.*, 2010). This further supports our crossing-out letters task as a depleting task.

Some limitations may be present in our dependent measures. Although our use of solvable anagrams is consistent with the depletion literature, our study would have benefitted from calibration of anagram difficulty. Cunningham and Baumeister (2016) observed that failed depletion replication studies are less likely to calibrate task difficulty with specific participant populations, such that tasks may be too easy (e.g., in an Ivy League sample) or too difficult (e.g., at a less selective campus), such that ceiling or floor effects may obscure depletion effects. In the current studies, the pooled mean number of solved anagrams, out of 30 possible, was less than 7 for both conditions (Control $M = 6.7$, Depletion $M = 6.3$), suggesting that the anagram difficulty level may have been high for this sample. Thus, it is possible that a depletion effect may have been diminished by the absence of task difficulty calibration.

Our grip duration dependent measure differs from most depletion studies in relying solely on post-manipulation handgrip duration rather than including a baseline handgrip duration for comparison. However, studies that measure baseline handgrip duration introduce a serious methodological confound. Specifically, engaging in a baseline handgrip duration task by definition is a depleting task, requiring exertion of self-control to hold the handgrip closed for as long as possible. Using such a procedure, the researcher does not have a non-depleted 'control' group to compare with a depleted group; instead, both groups are depleted, with the 'control' group depleted by the baseline grip duration task. While there is significant individual variability in handgrip duration that may reduce the ability to detect a depletion effect, we would argue that individual variability is preferable to the significant confound of a depleting baseline handgrip duration task in an ostensible control group. Accordingly, our procedure allows for a more straightforward test of depletion effects than studies that measure baseline handgrip duration.

Finally, our multiplication task in Studies 3A–3D likely serves as the weakest test of the depletion effect. As noted previously, some researchers do not view multiplication problems as requiring self-control (e.g., Muraven & Slessareva, 2003; Muraven *et al.*, 1998). At the same time, it has been argued that tasks vary in the extent to which self-control is required and that even lower complexity tasks, such as mental multiplication, demand a degree of self-control resources in the form of overriding the urge to discontinue, as well as engaging executive functions via maintaining and updating working memory (Hagger *et al.*, 2010). As such, while our multiplication task was not a strong dependent measure for detecting depletion effects, we can nonetheless predict a degree of performance deficit from the principles of resource depletion theory. Indeed, previous studies have in fact observed poorer multiplication task performance after an earlier depleting task (Wright *et al.*, 2008).

Although we conducted and presented these results as a series of small studies with sample sizes similar to those in the pre-2013 published literature ($n = 35$ per condition), statistical power was significantly increased by collapsing across the four studies within each of the three separate methods, with sample size increased to $n = 140$ per condition. Analyses at the pooled level also indicated no significant depletion effect for each task pairing. It could be argued that a true depletion effect was not detected because tasks were inadequately calibrated for difficulty (e.g., anagrams), or because tasks were not sufficiently demanding of self-control to detect a depletion effect (e.g., multiplication problems). Further, it is notable that the direction of the pooled non-significant effects in two of the three task pairings was in the direction of a depletion effect (i.e., means for the Control condition were above those of the Depletion condition). In addition, the overall pooled effect size across all participants in all studies was in the direction consistent with a depletion effect, although non-significantly so.

In summary, our studies did not indicate a robust depletion effect across multiple small studies, or with the higher statistical power afforded by pooled data. At the same time, our studies may not have provided the strongest possible tests of depletion effects in that some tasks may not have been adequately demanding of self-control resources, or may have required difficulty calibration for the local population. In addition, we observed that condition mean differences on the dependent measure were more often in the direction of a depletion effect than the reverse, even though statistical significance was not observed for any of the pooled data.

Our results suggest that the depletion effect is neither as robust nor as ubiquitous as suggested by the extant literature prior to the recent wave of challenges (e.g., Carter & McCullough, 2013, 2014; Lurquin *et al.*, 2016). At the same time, numerous earlier studies suggest that there may be some form of performance decrement attributable to prior engagement on a difficult task. It is likely that the general lack of consensus regarding the specific task types, duration, and intensity under which such an effect should occur has hampered a clear characterization of the hypothesized depletion effect. As such, a clear specification of tasks that are regarding as engaging self-control, along with a conceptual and empirical justification for identifying such tasks, would be valuable for advancing the field.

It is also likely that publication bias has obscured an understanding of the nature and parameters of a depletion effect, although there has been significant disagreement regarding criteria for inclusion and exclusion in depletion meta-analyses (e.g., Carter & McCullough, 2014; Cunningham & Baumeister, 2016). Recent proposals that researchers preregister their intended experimental protocols (e.g., Hagger *et al.*, 2015) would be of significant benefit in addressing this problem.

Disclosure

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Alberts, H. J., Martijn, C., Greb, J., Merckelbach, H., & de Vries, N. K. (2007). Carrying on or giving in: the role of automatic processes in overcoming ego depletion. *British Journal of Social Psychology*, *46*, 383–399. https://doi.org/10.1348/014466606x130111

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

Baumeister, R. F., & Heatherton, T. F. (1996). Self-regulation failure: An overview. *Psychological Inquiry*, *7*, 1–15. https://doi.org/10.1207/s15327965pli0701_1

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, *16*, 351–355. https://doi.org/10.1111/j.1467-8721.2007.00534.x

Bray, S. R., Ginis, K. A., Hicks, A. L., & Woodgate, J. (2008). Effects of self-regulatory strength depletion on muscular performance and EMG activation. *Psychophysiology*, *45*, 337–343. https://doi.org/10.1111/j.1469-8986.2007.00625.x

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*, 796–815. https://doi.org/10.1037/xge0000083

Carter, E. C., & McCullough, E. (2013). Is ego depletion too incredible? Evidence for the overestimation of the depletion effect *Behavioral and Brain Sciences*, *36*, 683–684. DOI: https://doi.org/10.1017/s0140525x13000952

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

Cunningham, M. R., & Baumeister, R. F. (2016). How to make nothing out of something: Analyses of the impact of study sampling and statistical interpretation in misleading meta-analytic conclusions. *Frontiers in Psychology*, *7*, 1639. https://doi.org/10.3389/fpsyg.2016.01639

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 Bulleting*, *34*, 1653–1662. https://doi.org/10.1177/0146167208323981

Dvorak, R. D., & Simons, J. S. (2009). Moderation of resource depletion in the self-control strength model: Differing effects of two modes of self-control. *Personality and Social Psychology Bulletin*, *35*, 572–583. https://doi.org/10.1177/0146167208330855

Fennis, B. M., Janssen, L., & Vohs, K. D. (2009). Acts of benevolence: A limited resource account of compliance with charitable requests. *Journal of Consumer Research*, *35*, 906–924. https://doi.org/10.1086/593291

Finkel, E. J., & Campbell, W. K. (2001). Self-control and accommodation in close relationships: An interdependence analysis. *Journal of Personality and Social Psychology*, *81*, 263–277. https://doi.org/10.1037//0022-3514.81.2.263

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. https://doi.org/10.1037/0022-3514.92.2.325

Gailliot, M. T., Peruche, B. M., Plant, E. A., & Baumeister, R. F. (2009). Stereotypes and prejudice in the blood: Sucrose drinks reduce prejudice and stereotyping. *Journal of Experimental Social Psychology*, *45*, 288–290. https://doi.org/10.1016/j.jesp.2008.09.003

Gailliot, M. T., Plant, E. A., Butz, D. A., & Baumeister, R. F. (2007). Increasing self-regulatory strength can reduce the depleting effect of suppressing stereotypes. *Personality and Social Psychology Bulletin*, *33*, 281–294. https://doi.org/10.1177/0146167206296101

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., . . . Zwienenberg, M. (2015). A multi-lab pre-registered replication of the ego- depletion effect. *Perspectives on Psychological Science*, *11*, 546–573. https://doi.org/10.1177/1745691616652873

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. https://doi.org/10.1037/a0019486

Heatherton, T. F., & Baumeister, R. F. (1996). Self-regulation failure: Past, present, and future. *Psychological Inquiry*, *7*, 90–98. https://doi.org/10.1207/s15327965pli0701_20

Inzlicht, M., McKay, L., & Aronson, J. (2006). Stigma as ego depletion: How being the target of prejudice affects self-control. *Psychological Science*, *17*, 262–269. https://doi.org/10.1111%2fj.1467-9280.2006.01695.x

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, *7*, 1–8. https://doi.org/10.7771/1932-6246.1167

Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General*, *137*, 691–705. https://doi.org/10.1037%2fa0013834

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lange, F., & Eggert, F. (2014). Sweet delusion: Glucose drinks fail to counteract ego depletion. *Appetite*, *75*, 54–63. https://doi.org/10.1016/j.appet.2013.12.020

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., von Bastian, C. C., Carruth, N. P., & Miyake, A. (2016). No evidence of the ego-depletion effect across task characteristics and individual differences: A pre-registered study. *PLoS ONE*, *11*, e0147770. https://doi.org/10.1371/journal.pone.0147770

Masicampo, E. J., & Baumeister, R. F. (2008). Toward a physiology of dual-process reasoning and judgment: Lemonade, willpower, and expensive rule-based analysis. *Psychological Science*, *19*, 255–260. https://doi.org/10.1111%2fj.1467-9280.2008.02077.x

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009). Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of Experimental Social Psychology*, *45*, 594–597.

Muraven, M. (2010). Building self-control strength: Practicing self-control leads to improved self-control performance. *Journal of Experimental Social Psychology*, *46*, 465–468. https://doi.org/10.1016%2fj.jesp.2009.12.011

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126*, 247–259. https://doi.org/10.1037/0033-2909.126.2.247

Muraven, M., & Slessareva, E. (2003). Mechanisms of self-control failure: Motivation and limited resources. *Personality and Social Psychology Bulletin*, *29*, 894–906.

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as a limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology*, *74*, 774–789. http://psycnet.apa.org/doi/10.1037/0022-3514.74.3.774

Murtagh, M., & Todd, S. A. (2004). Self-regulation: a challenge to the strength model. *Journal of Articles in Support of the Null Hypothesis*, *3*, 19–50.

Park, S. H., Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintended discrimination. *Social Cognition*, *26*, 401–419. https://doi.org/10.1521/soco.2008.26.4.401

R Core Team. (2014). *R: A language and environment for statistical computing (version 3.4.2) [Computer software]*. Vienna, Austria: R Foundation for Statistical Computing.

Roever, C., & Friede, T. (2017). R package bayesmeta version 1.6 [Computer software]. Retrieved from https://cran.r-project.org/web/packages/bayesmeta/index.html

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566. https://doi.org/10.1037/a0029487

Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *Journal of Experimental Psychology: General*, *136*, 241–255. https://doi.org/10.1037/0096-3445.136.2.241

Tyler, J. M., & Burns, K. C. (2008). After depletion: The replenishment of the self's regulatory resources. *Self and Identity*, *7*, 305–321. https://doi.org/10.1080/15298860701799997

Vohs, K. D., Baumeister, R. F., & Ciarocco, N. J. (2005). Self-regulation and self-presentation: Regulatory resource depletion impairs impression management and effortful self-presentation depletes regulatory resources. *Journal of Personality and Social Psychology*, *88*, 632–657. https://doi.org/10.1037/0022-3514.88.4.632

Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *Psychological Science*, *11*, 249–254. https://doi.org/10.1111%2f1467-9280.00250

Wallace, H. M., & Baumeister, R. F. (2002). The effects of success versus failure feedback on further self-control. *Self and Identity*, *1*, 35–41.

Wan, E. W., & Sternthal, B. (2008). Regulating the effects of depletion through monitoring. *Personality and Social Psychology Bulletin*, *34*, 32–46. https://doi.org/10.1177/0146167207306756

Wright, R. A., Junious, T. R., Neal, C., Avello, A., Graham, C., Herrmann, L., . . . Walton, N. (2007). Mental fatigue influence on effort-related cardiovascular response: difficulty effects and extension across cognitive performance domains. *Motivation and Emotion*, *31*, 219–231. https://doi.org/10.1007/s11031-007-9066-9

Wright, R. A., Stewart, C. C., & Barnett, B. R. (2008). Mental fatigue influence on effort-related cardiovascular response: Extension across the regulatory (inhibitory)/non-regulatory performance dimension. *International Journal of Psychophysiology*, *69*, 127–133. https://doi.org/10.1016/j.ijpsycho.2008.04.002

Received 26 April 2017; revised version received 8 December 2017