Habituation Does Not Rescue Depletion: Two Tests of the Ego-Depletion Effect

Author Note

Anonymous version for refereeing. This Version: October 2018

Abstract

The recent literature on ego depletion and self-control is plagued with failed replications. It has been argued that publication bias might have inflated estimated effect sizes. Doubts go so far that the very existence of the ego-depletion effect has been questioned. We conducted two high-power tests of the ego-depletion effect, with samples in two different countries, including a habituation phase in the depleting task (“e-crossing task”). This addresses recent critiques on failed registered replications, which argued that habituation was essential to obtain depletion effects. We examined the effect on error rates, response times, and response-time variability in a subsequent Stroop task. There were no effects in general, except a significant difference in response-time variability, only in one of the samples, and in the opposite direction as predicted by the ego-depletion effect.

Keywords: ego depletion, self-control

Habituation Does Not Rescue Depletion: Two Tests of the Ego-Depletion Effect

The ego-depletion effect (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Muraven, Tice, & Baumeister, 1998), where exercising self-control in a task results in diminished self-control in subsequent tasks, has been discussed for over 20 years. A meta-analysis by Hagger, Wood, Stiff, and Chatzisarantis (2010) revealed a medium-sized effect ($d = 0.62$) across 198 studies, but Carter and McCullough (2013) argued that publication bias could have led to an overestimation of the effect. Further discussion (Carter & McCullough, 2014; Hagger & Chatzisarantis, 2014) culminated in a registered replication report (Hagger et al., 2016), which failed to find significant effects. This has led to skepticism and thorough questioning of the evidence (see, e.g., Gissubel, Beiramar, & Freire, 2018).

The results of Hagger et al. (2016) were contested by Baumeister and Vohs (2016) due to the implementation of the initial task. In this task, based on Baumeister et al. (1998), self-control resources are depleted by crossing out instances of the letter "e" according to a complex rule with several exceptions (compared with the control, where all instances are crossed out). Inhibiting the impulse to cross out the letter requires self-control. Hagger et al. (2016) used a computerized version of the task by Sripada, Kessler, and Jonides (2014) which did not include a so-called habituation phase. Baumeister and Vohs (2016) argued that the habituation phase is necessary to create the habit which is to be inhibited later. In contrast, Arber et al. (2017) argued that an e-crossing task even without habituation phase includes several steps that require self-control, and thus should also be depleting. It should be noted that Baumeister and Vohs (2016) also criticized the computerized version. However, that format was also successfully used in Alós-Ferrer, Hügelschäfer, and Li (2015), and Arber et al. (2017) showed depletion effects for both versions (computerized and paper-and-pencil) of the e-crossing task.

Following Baumeister and Vohs (2016), an ego-depletion effect should obtain if a habituation phase is included. However, the ego-depletion effect might be weak. In view of the failed registered replication of Hagger et al. (2016), Friese, Loschelder, Gieseler, Frankenbach, and Inzlicht (2018) analyze various arguments for and against the evidence of the existence of an ego-depletion effect. The mai nargument against it is that possible publication bias might have lead to an overestimation of the effect size, and hence the ego-depletion effect might be weak or even non-existent. A third view is that ego-depletion exists but is moderated by the susceptibility to lay theories of limited willpower. Job, Dweck, and Walton (2010) show ego-depletion effects for subjects who hold beliefs about limited willpower, while subjects believing in unlimited willpower did not show diminished self-control.

In two separate studies (carried out in different countries), we tested the ego-depletion effect adding a habituation phase to the task of Sripada et al. (2014), with large sample sizes determined to achieve more than sufficient power. Our hypothesis was that the ego-depletion effect would be observed and moderated by implicit theories of self-control as argued by Job et al. (2010). That is, we allowed for the possibility that ego-depletion effects are only relevant if the participant's lay views on self-control endorse a limited-resource view.

Method

To determine sample size, we followed Hagger et al. (2016), who conducted an ex ante power analysis with $\alpha = 0.01$ and $1 - \beta = 0.95$ for a medium effect size ($d = 0.62$) based on the meta-analysis of Hagger et al. (2010). With these assumptions, they calculated 168 participants (84 in each condition) for a one-sided test. Given the commonly-applied (and demanding) inclusion criteria (Karalunas, Huang-Pollock, & Nigg, 2013; Sripada et al., 2014 excluded 14 out of 108 subjects, i.e. 12.96%), we increased sample size and rounded up taking into account lab sizes. In Experiment 1, we conducted 8 sessions of at most 30 participants each. Taking into account no-shows, we collected data for $N = 229$ participants. In Experiment 2, we conducted 5 sessions of 48 participants each and collected data for $N = 240$ participants (there were no no-shows).[1] We relied on Hagger et al. (2016) for the power analysis for comparability. However, considering that the effect size might be overestimated, we remark that, even for a small-to-moderate effect size ($d = 0.35$), a sample of $N = 215$ and a probability level $\alpha = 0.05$ still yields a power of 0.82.

Both experiments followed the same design. First, participants completed a number of scales including the 6-items scale of Job et al. (2010) to assess implicit theories of self-control ("lay self-control"). It contains six items with answers ranging from 1 (strongly agree) to 6 (strongly disagree). In Experiment 1, carried out in Germany, we used the original items of Job et al. (2010).[2] For Experiment 2, we translated the items into Spanish using a back-translation procedure involving two independent translators. Following Job et al. (2010), the scale was embedded among several other implicit-theory measures, to avoid making the purpose of the scale obvious. The exact order of the scales was: implicit theories of intelligence, willpower, resisting temptations, character, and emotions. Then, participants conducted the "e" task as in Sripada et al. (2014) and Alós-Ferrer et al. (2015), with the exception that a habituation phase was added. Specifically, for 32 trials, all participants had to respond by pressing a key if a word displayed on screen contained the letter "e." Then, 68 further trials followed, where controls performed the exact same task but participants assigned to the depletion condition had to press the key if the word contained an "e," *unless* a vowel followed the "e" *or* there was another vowel two letters away from the "e" in *either* direction.[3] A single trial lasted 3000 ms and the whole "e" task lasted 5 minutes. The sets of words were chosen to have identical target rates in both experiments, independently of the language. After the "e" task was completed, participants were asked to indicate how exhausting they found the task.

[1] Sample size should also be sufficient to detect possible interactions with implicit theories of self-control. According to a G*Power calculation, an effective sample size of $N = 200$ would give a power of 0.82 for detecting a medium effect size ($f = 0.25$) at $\alpha = 0.01$ through the interaction term in an ANCOVA.

[2] We thank Veronika Job for facilitating the original scale and advising us on how to implement it.

[3] Due to a programming error, in Experiment 2 the second phase consisted of 67 trials instead of 68.

After the "e" task, participants completed a computerized Stroop task for 72 trials: 24 congruent words naming a color printed in the same color, 24 incongruent words naming a different color to the one the word was printed in, and 24 neutral words printed in some color but naming a non-color word. Four different colors were included. A single trial lasted 3000 ms and for the analysis a few trials (0.01%) which had extremely short response times (below 250 ms) were recoded as missing. The different types of trials were interleaved. The hypothesis was that depleted participants would have less resources to inhibit their responses in incongruent trials. After the Stroop task, subjects were again asked to indicate how exhausting they found the task. Finally, participants filled in the 13-item Brief Self-Control Scale of Tangney, Baumeister, and Boone (2004) and an exploratory scale (not used here), and were requested to provide demographic data (e.g., age and gender). The German version of the Brief Self-Control Scale was taken from Bertrams and Dickäuser (2009); since no Spanish version was available in the literature at that point, we translated the items into Spanish using a back-translation procedure involving two independent translators.

We relied on three measures of performance in the Stroop task. The first is the percentage of errors for which some studies find depleted participants to have higher error rates in incongruent trials compared with controls (Friese, Binder, Luechinger, Boesiger, & Rasch, 2013; Gailliot, Schmeichel, & Baumeister, 2006). However, error rates in the Stroop task are typically very small, and hence many studies focus on other measures (Fennis, Janssen, & Vohs, 2008; Richeson & Trawalter, 2005). We hence did not wish to confine ourselves to this measure. The second is response times, where the prediction would be longer response times in incongruent trials for depleted participants compared with controls (Fennis et al., 2008; Muraven, Shmueli, & Burkley, 2006). For instance, Inzlicht and Gutsell (2007) find an effect in response times in the Stroop task even though there was no effect on error rates. For both errors and response times, as an additional performance measure we also rely on the interference between incongruent and congruent trials (that is, the difference between error rates or response times between both types of trials) as well as between incongruent and neutral trials. The final performance measure is response time variability (RTV) following Sripada et al. (2014) and Hagger et al. (2016). RTV is calculated by adding the σ and τ parameters of a fitted ex-Gaussian distribution of response times for each individual. The ex-Gaussian distribution is derived from the sum of a Gaussian (determined by μ and σ) and an exponential distribution (determined by τ). The characteristics of the ex-Gaussian distribution fit response time data very well (Dawson, 1988) and the sum of the variability parameters σ and τ, i.e. RTV, is considered a correlate of attentional control (see Sripada et al., 2014, for details). The hypothesis is that depleted participants would exhibit higher RTV in incongruent trials, reflecting a difficulty in maintaining task-directed focus.

Both experiments were programmed in z-Tree (Fischbacher, 2007). Ex-gaussian modeling was programmed in R and data were analyzed using Stata. For response time analyses, we followed the procedures for exclusions described in Karalunas et al. (2013) and Sripada et al. (2014). Specifically, for each participant we excluded observations in the Stroop task with response times two standard deviations away from the individual mean for correct responses in incongruent trials, and analogously for congruent and neutral trials. We also had to exclude a colour-blind participant and 7 further participants due to technical problems. After all exclusions, the sample size was $N = 225$ in Experiment 1 and $N = 236$ in Experiment 2. For the RTV analyses (and only for those) we further excluded participants with RTVs two standard deviations away from the experiment's mean (15 exclusions), yielding $N = 222$ and $N = 224$ for Experiment 1 and Experiment 2, respectively.

The registered replication of Hagger et al. (2016) used a demanding exclusion rule by which all participants with an error rate of 20% or more in the "e" task were excluded from the analysis. This criterion is questionable, because if the manipulation works, participants in the depletion condition face a cognitively more demanding task than controls, and hence it is to be expected that the exclusion rule leads to a selection problem where significantly more participants are excluded in the depletion condition than in the control condition. Higher error rates for the depletion task are, simply put, part of the nature of the task. This was overwhelmingly the case in both of our (independent) samples. Surprisingly, Hagger et al. (2016) report that this only occurred in 5 of the 23 studies they include, raising reasonable doubts as to whether their manipulation (without habituation) actually successfully induced ego depletion. In view of this possible criticism, we analyzed our data both with the exclusion rule mentioned above and without it. However, we obtained exactly the same results in both cases, and hence report here the largest sample, adding only a summary of the analysis of the restricted sample at the end of the Results section.

For each experiment, we conducted t-tests for response times, errors, and RTVs. We conducted ANCOVAs taking advantage of the full range of the lay self-control scale. As an additional illustration, we further implemented a median split of participants with subjects scoring higher on the lay self-control scale being those with a higher belief in willpower as a limited resource. We then also conducted the key tests separately for the two resulting subsamples.

Results

Experiment 1

Experiment 1 was conducted at a large German university. After exclusions, there were $N = 225$ participants (120 females; age 17 to 33, $M = 21.59$, $SD = 2.78$), of which 115 were controls and 110 were in the depletion condition. The depletion version of the "e" task was harder as reflected by response times: the mean RTs were 1733 ms (95% CI= $[1683, 1784]$) in the depletion condition and 753 ms (95% CI= $[725, 781]$) for controls, $t(223) = 33.97$, $p < .001$, Cohen's $d = 4.53$. The participants also differed in the error rates of the second phase of the depletion task. The error rates were $M = 10.76\%$ for depleted participants (95% CI= $[8.52, 13.00]$) and $M = 0.79\%$ for controls (95% CI= $[0.04, 1.55]$).

The difference was highly significant, $t(223) = 8.50$, $p < .001$, $d = 1.13$. A non-parametric Wilcoxon Rank-Sum test (more appropriate for a comparison of error rates, which are bound to the $[0, 1]$ interval) also revealed a highly significant difference, $z = 11.99$, $p < .001$. These measurements were in line with the self-reported exhaustion level after the task. On a scale from 0 (not exhausting at all) to 100 (very exhausting), participants in the depletion group reported an average of 44.63 (95% CI= $[39.64, 49.61]$) which was significantly higher than the average of 11.79 (95% CI= $[8.70, 14.88]$) reported by the control group ($t(223) = 11.21$, $p < .001$, Cohen's $d = 1.49$). The lay self-control scale was reliable ($N = 225$, Cronbach's $\alpha = .84$, $M = 4.08$, $SD = 0.81$).

In the Stroop task, error rates for correct responses in incongruent trials were very low, $M = 3.02\%$ for depleted participants (95% CI= $[2.15, 3.89]$) and $M = 3.64\%$ for controls (95% CI= $[2.73, 4.55]$). The difference was not significant, $t(223) = -0.97$, $p = 0.333$, $d = -0.13$. A non-parametric Wilcoxon Rank-Sum test also revealed no significant difference, $z = -1.00$, $p = .316$.

Results for response times and RTVs are illustrated in Figure 1. Our first measure are response times for correct responses in incongruent trials of the Stroop task. Those were not significantly different for depleted participants ($M = 1170$ ms, 95% CI= $[1137, 1202]$) and controls ($M = 1175$ ms, 95% CI= $[1143, 1208]$), $t(223) = -0.23$, $p = .818$, $d = -0.03$. A one-way ANCOVA also revealed no effect of ego depletion on mean response times controlling for the scores on implicit theory of willpower, $F(1, 222) = 0.05$, $p = .817$, $\eta^2 = 0.000$. The differences in response times were also not significant for participants endorsing an unlimited theory (depleted: $N = 51$, $M = 1163$ ms, 95% CI= $[1115, 1211]$; controls: $N = 54$, $M = 1155$ ms, 95% CI= $[1105, 1206]$; $t(103) = 0.22$, $p = .828$, $d = 0.04$) or those endorsing a limited-resource theory of self-control (depleted: $N = 59$, $M = 1176$ ms, 95% CI= $[1130, 1221]$; controls: $N = 61$, $M = 1193$ ms, 95% CI= $[1151, 1235]$; $t(118) = -0.55$, $p = .585$, $d = -0.10$).

The same conclusions were obtained using the Stroop interference scores for errors and response times between depleted participants and the control group. The average interference score for errors between incongruent trials and congruent trials was $M = 1.80\%$ for depleted participants (95% CI= $[0.86, 2.74]$) and $M = 1.95\%$ for controls (95% CI= $[0.93, 2.96]$). The difference was not significant ($t(223) = -0.21$, $p = 0.837$, $d = -0.03$). The average interference score for response times was $M = 113$ ms for depleted participants (95% CI= $[93, 132]$) and $M = 102$ ms for controls (95% CI= $[83, 121]$). The difference was also not significant ($t(223) = 0.79$, $p = 0.432$, $d = 0.11$). A one-way ANCOVA also revealed no effect of ego depletion on mean interference scores for response times controlling for implicit theories of willpower ($F(1, 222) = 0.61$, $p = .434$, $\eta^2 = 0.003$). Individual tests in each subsample created by a median split with respect to willpower theories also failed to reveal any significant difference.[4]

Our third measure are RTVs as described above. Again, there were no significant differences between depleted participants ($M = 0.31$, 95% CI= $[0.30, 0.33]$) and controls ($M = 0.32$, 95% CI= $[0.30, 0.34]$), $t(220) = -0.33$, $p = .742$, $d = -0.04$. A one-way ANCOVA also revealed no effect of ego depletion on RTVs controlling for the scores on implicit theory of willpower, $F(1, 219) = 0.13$, $p = .723$, $\eta^2 = 0.001$. The result remains unchanged for unlimited-theory participants (depleted, $N = 50$, $M = 0.32$, 95% CI= $[0.29, 0.35]$; controls, $N = 53$, $M = 0.32$, 95% CI= $[0.30, 0.35]$; $t(101) = -0.21$, $p = .834$, $d = -0.04$) and limited-theory ones (depleted, $N = 58$, $M = 0.31$, 95% CI= $[0.29, 0.34]$; controls, $N = 61$, $M = 0.32$, 95% CI= $[0.29, 0.34]$; $t(117) = -0.25$, $p = .801$, $d = -0.05$).

Experiment 2

Experiment 2 was conducted at a large Spanish university. After exclusions, there were $N = 236$ participants (132 females; age 17 to 37, $M = 20.22$, $SD = 2.39$), of which 118 were controls and 118 were in the depletion condition. The depletion version of the "e"

[4] We also tested for differences in the Stroop interference scores between incongruent and neutral trials. There were no significant differences.

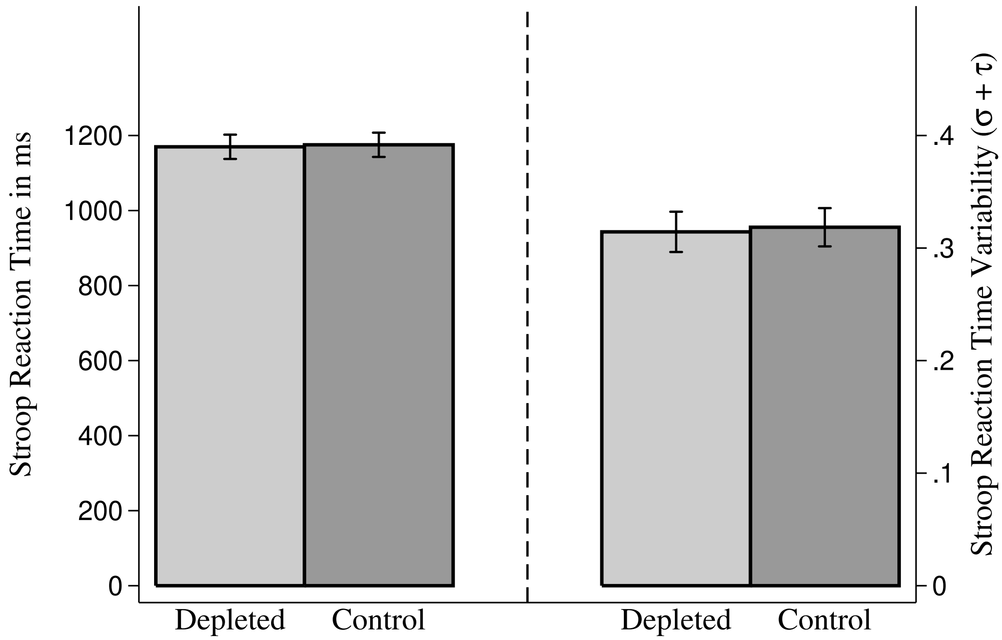

Figure 1. Experiment 1. Comparison of response times and RTVs between depleted participants and controls.

task was harder as reflected by response times, error rates, and self-reported exhaustion levels. The mean response times for correct responses were 1588 ms (95% CI= $[1537, 1639]$) in the depletion condition[5] and 700 ms (95% CI= $[679, 720]$) for controls ($t(233) = 32.03$, $p < .001$), Cohen's $d = 4.18$. The mean error rates in the second phase of the depletion task were $M = 11.35\%$ for depleted participants (95% CI= $[9.12, 13.57]$) and $M = 0.39\%$ for controls (95% CI= $[0.24, 0.55]$). The difference was highly significant ($t(234) = 9.72$, $p < .001$, $d = 1.27$). A non-parametric Wilcoxon Rank-Sum test also revealed a highly significant difference, $z = 12.93$, $p < .001$. These measurements were in line with the self-reported exhaustion level after the task. On a scale from 0 (not exhausting at all) to 100 (very exhausting), participants in the depletion group reported an average of 36.89 (95% CI= $[32.12, 41.66]$) which was significantly higher than the reported average of 21.32

[5]One subject failed to provide any correct response.

(95% CI$= [17.01, 25.64]$) of the control group ($t(234) = 4.79$, $p < .001$, Cohen's $d = 0.62$). The lay self-control scale was reliable ($N = 236$, Cronbach's $\alpha = .74$, $M = 4.87$, $SD = 0.77$).

Error rates in the Stroop task were not significantly different between depleted participants ($M = 2.88\%$, 95% CI$= [2.15, 3.61]$) and controls ($M = 2.56\%$, 95% CI$= [1.84, 3.28]$), $t(234) = 0.62$, $p = 0.536$, $d = 0.08$. A non-parametric Wilcoxon Rank-Sum test also revealed no significant difference, $z = 0.81$, $p = .418$.

Results for response times and RTVs are illustrated in Figure 2. As in Experiment 1, response times for correct responses in incongruent trials of the Stroop task were not significantly different for depleted participants ($M = 1161$ ms, 95% CI$= [1130, 1192]$) and controls ($M = 1130$ ms, 95% CI$= [1104, 1156]$), $t(234) = 1.51$, $p = .132$, $d = 0.20$. A one-way ANCOVA also revealed no effect of ego depletion on mean response times controlling for the scores on implicit theory of willpower, $F(1, 233) = 2.49$, $p = .116$, $\eta^2 = 0.011$. The differences in response times were also not significant for participants endorsing an unlimited-resource theory of self-control (depleted: $N = 55$, $M = 1145$ ms, 95% CI$= [1103, 1188]$; controls: $N = 52$, $M = 1115$ ms, 95% CI$= [1073, 1156]$; $t(105) = 1.04$, $p = .302$, $d = 0.20$), or for those endorsing a limited theory (depleted: $N = 63$, $M = 1175$ ms, 95% CI$= [1130, 1220]$; controls: $N = 66$, $M = 1143$ ms, 95% CI$= [1109, 1176]$; $t(127) = 1.15$, $p = .253$, $d = 0.20$).

The same conclusions were obtained using Stroop interference scores for errors and response times between depleted participants and the control group. The average interference score for errors between incongruent and congruent trials was $M = 1.29\%$ for depleted participants (95% CI$= [0.46, 2.12]$) and $M = 1.22\%$ for controls (95% CI$= [0.44, 2.00]$). The difference was not significant, $t(234) = 0.11$, $p = 0.910$, $d = 0.01$. The average interference score for response times was $M = 105$ ms for depleted participants (95% CI$= [86, 123]$) and $M = 98$ ms for controls (95% CI$= [80, 116]$). The difference was also not significant, $t(234) = 0.50$, $p = 0.619$, $d = 0.06$. A one-way ANCOVA also failed to reveal any effect of ego depletion on mean interference scores for response

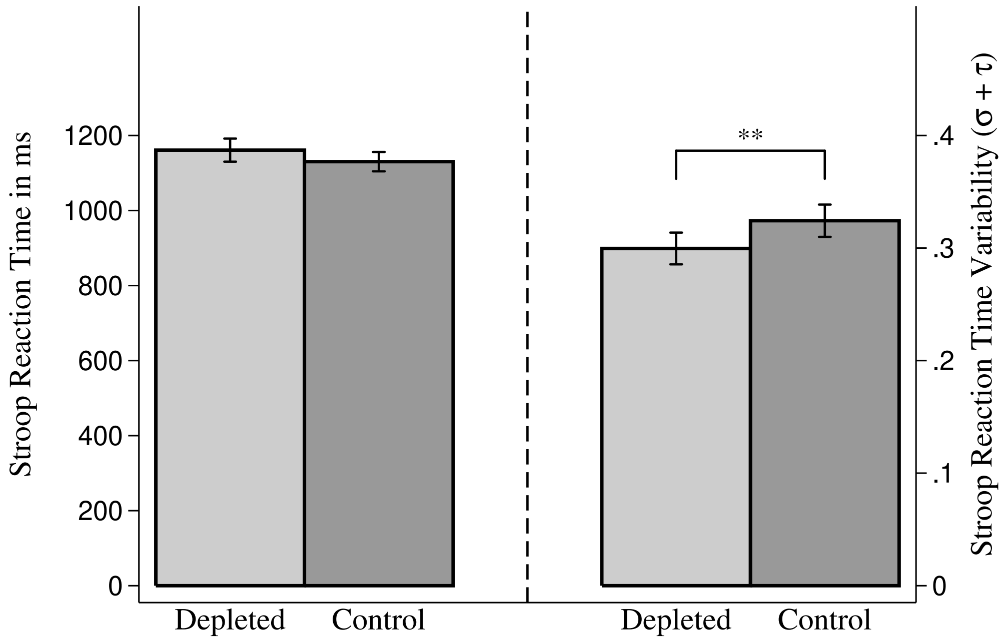

Figure 2. Experiment 2. Comparison of response times and RTVs between depleted participants and controls.

times controlling for implicit theories of willpower ($F(1, 233) = 0.25$, $p = .620$, $\eta^2 = 0.001$). Individual tests in each subsample created by a median split according to willpower theories also did not reveal any significant differences.[6]

In contrast to Experiment 1, the difference in RTVs was clearly significant, but in the opposite direction as expected. RTVs for depleted participants ($M = 0.30$, 95% CI$= [0.29, 0.31]$) were significantly lower than RTVs for controls ($M = 0.32$, 95% CI$= [0.31, 0.34]$), $t(222) = -2.43$, $p = .016$, Cohen's $d = -0.32$. A one-way ANCOVA confirmed a significant effect of ego depletion on RTVs after controlling for the scores on implicit theory of willpower, $F(1, 221) = 5.90$, $p = .016$, $\eta^2 = 0.026$. The result comes from participants with an unlimited-theory of willpower (depleted, $N = 51$, $M = 0.30$, 95%

[6] As in Experiment 1, there were also no differences in the Stroop interference scores for incongruent vs. neutral trials.

CI= $[0.28, 0.32]$; controls, $N = 49$, $M = 0.34$, 95% CI= $[0.32, 0.36]$; $t(98) = -2.80$, $p = .006$, $d = -0.56$); for limited-theory participants, RTVs for depleted participants ($N = 61$, $M = 0.30$, 95% CI= $[0.28, 0.32]$) were not significantly lower than RTVs for controls ($N = 63$, $M = 0.31$, 95% CI= $[0.29, 0.33]$), $t(122) = -0.74$, $p = .463$, $d = -0.13$).

Alternative exclusion rule

We repeated the whole analysis with the exclusion rule used by Hagger et al. (2016). That is, we obtained reduced samples by first excluding participants with error rates above 20% in the "e" task. This led to 36 exclusions, of which 35 were in the depletion treatments and only one in the control treatments. Then, we applied the previous exclusion criteria for response-time analysis, that is, for each participant we excluded observations in the Stroop task with response times two standard deviations away from the individual mean (for correct responses in incongruent trials). As in the full-sample case, we excluded a colour-blind participant and 7 further participants due to technical problems. After all exclusions, the reduced sample size was $N = 209$ in Experiment 1 and $N = 216$ in Experiment 2. Of course, exclusions happened significantly more often for the depletion groups (Experiment 1: $\chi^2(1) = 11.42$, $p = .001$; Experiment 2: $\chi^2(1) = 18.52$, $p < .001$). For the analysis of RTVs we then excluded participants with RTVs two standard deviations away from the reduced sample's mean (14 exclusions). The final sample size for the analysis of RTV is $N = 205$ in Experiment 1 and $N = 206$ in Experiment 2.

All tests and analyses were qualitatively unchanged with the reduced samples with one exception. In Experiment 1, a one-way ANCOVA confirmed a significant effect of ego depletion on Stroop-task error rates after controlling for the scores on implicit theory of willpower, $F(1, 206) = 2.90$, $p = .090$, $\eta^2 = 0.014$. A t-test confirmed that average error rates in the Stroop task in Experiment 1 were significantly smaller for the ego-depletion group ($M = 2.57\%$, 95% CI= $[1.74, 3.39]$) than for the control group ($M = 3.67\%$, 95% CI= $[2.75, 4.59]$). This is weakly significant ($t(207) = -1.74$, $p = .083$, $d = -0.24$), but unsurprising since the alternative exclusion rule led to the exclusion of high error rates which happened more often in the ego-depletion group. In contrast, the more appropriate non-parametric Wilcoxon Rank-Sum test shows no significant difference ($z = -1.49$, $p = .136$). Other results are unchanged and not significant, except for those for RTVs in Experiment 2, which were as in the full sample.

Exploratory Analyses

We included the brief version of the Trait Self-Control Scale of Tangney et al. (2004), which was reliable (Experiment 1: $N = 225$, $\alpha = 0.80$, $M = 41.18$, $SD = 7.03$; Experiment 2: $N = 236$, $\alpha = 0.78$, $M = 41.08$, $SD = 6.96$). The scale did not significantly correlate with error rates, response times, or RTVs in the Stroop task in either experiment. There were also no significant correlations when splitting the samples in depleted participants and controls, with only one exception: for depleted participants in Experiment 1, the scale correlated positively with RTVs ($N = 108$, $r = 0.17$, $p = .084$). This relation was not found for controls in Experiment 1 and was not replicated in Experiment 2, neither for depleted participants nor for controls.

Our ego-depletion task provides a quantifiable measure of performance in the form of error rates in the e-crossing task. In an exploratory analysis we compared the performance in the depletion task with various Stroop performance measures, separately for the ego-depletion and the control group. in both experiments, we find a significant positive correlation between errors in the e-crossing task and errors in the incongruent trials in the Stroop task for the depleted group (Experiment 1, $N = 110$, $r = 0.25$, $p = .010$; Experiment 2, $N = 118$, $r = 0.18$, $p = .053$). The correlations were not significant for the control group. This is unsurprising, because error rates in the e-crossing task are very low for the control group. The correlation for the depletion group simply reflects individual differences. We also found significant correlations with other measures, but not systematically across experiments.[7]

Discussion

Contrary to our hypotheses, we cannot confirm the predictions of the ego-depletion effect, independently of whether error rates, response times, interference scores for error rates, interference scores for response times, or response-time variability are used as dependent variables. In one experiment, but not in the other, there was a clear effect on response-time variability, which however went in the opposite direction as expected according to the interpretation of this variable (Karalunas et al., 2013; Sripada et al., 2014) and the ego-depletion effect.

Of course, we cannot conclude that the ego-depletion effect is a mirage, and this was not our intention when we set to carry out the studies reported here. However, our evidence suggests four possibilities. First, *inducing* ego depletion might be a subtler proposition than previously assumed. The "e" task was considered highly effective for this purpose (Hagger & Chatzisarantis, 2014). Despite positive results of its computerized version (Alós-Ferrer et al., 2015; Sripada et al., 2014), the registered replication of Hagger et al. (2016) failed to find effects using it, and Baumeister and Vohs (2016) criticized the lack of a habituation phase. We failed to find the expected results even though we used a habituation phase, hence this might not be the difficulty after all.

Second, *measuring* ego depletion, even if successfully induced, might be harder than usually assumed. According to the depletion logic, a common resource is employed in both

[7]In Experiment 1 (but not in Experiment 2), errors in the e-crossing task correlated with Stroop-error interference scores for both incongruent vs. congruent trials ($N = 110$, $r = 0.16$, $p = .088$) and incongruent vs. neutral trials ($N = 110$, $r = 0.24$, $p = .011$). In Experiment 2 (but not in Experiment 1), errors in the e-crossing task correlated with response times ($N = 118$, $r = 0.20$, $p = .028$) and RTV ($N = 112$, $r = 0.16$, $p = .085$) in the Stroop task. Errors in the e-crossing task did not correlate with interference scores for response times in the Stroop task in either experiment. No correlation of any type was observed for the control group in either experiment.

the manipulation task and the task used to measure the effects. The Stroop task is standard in ego-depletion studies, but our results might be evidence of an asymmetry, that is, tasks which work well to induce depletion might not be suitable to measure its effects.

Third, the *scope of the tasks* considered in the literature might be too broad. It might be necessary to consider a finer taxonomy of the cognitive functions which are temporarily weakened by specific ego-depletion manipulations. For instance, manipulations as the Stroop task require strong cognitive inhibition processes, while alternative manipulations involve rather different processes. An extreme example is the cold pressor task (Schmeichel & Vohs, 2009; Vohs et al., 2008), which relies on processes of pain control. Ego-depletion effects might simply be too weak if the manipulation and the subsequent task require different cognitive functions or processes. However, as just discussed, the obvious possibility to use the exact same task or very similar ones might not be a solution.

Last, the *consequences* of ego-depletion might be less well-understood than usually assumed. Our second experiment found an effect contrary to current interpretations of ego depletion, which was absent in our first experiment, conducted in a different country. Previous studies involving the same countries where we collected data (Achtziger, Alós-Ferrer, & Wagner, 2015, 2016, 2018) have found different effects of ego depletion across countries in the domain of social preferences. This is perhaps not so surprising, as it can be argued that there is a marked heterogeneity in what constitutes default behavior in that domain, possibly giving rise to country effects or cultural differences. Savani and Job (2017) found a *reverse* ego-depletion effect in India compared to the Western culture. Why there should be cultural differences between two Western countries (Germany and Spain) in effects and manipulations as simple as those studied here, however, remains an open question.

These possibilities are, of course, not mutually exclusive. Rather, they suggest possible avenues for a revision of the theories underlying ego-depletion effects, with the aim of explaining why those effects are sometimes observed and sometimes missed. Our results suggest that, in order to accomplish this objective, such a revision might need to incorporate previously neglected elements.

Our studies have of course limitations. First, it might be that, even including a habituation phase, the e-crossing task was not sufficiently depleting. Brown and Bray (2017) find that depletion effects occur after reaching a threshold somewhere between 4 and 6 minutes (for their depletion task). Similarly, Arber et al. (2017) found that, after a certain threshold, performance in the depletion task itself significantly decreased, which they interpreted as a sign of diminished self-control. The task was not comparable to ours, as it relied on certain stories (performance started decreasing between story two and three, out of five). We did not find a steady decrease in performance in the e-crossing task with the number of trials, in contrast to Arber et al. (2017). The duration of our e-crossing task was 5 minutes, which has been found to be effective in causing diminished self-control in other studies (Achtziger et al., 2018; Alós-Ferrer et al., 2015). However, our manipulation checks indicated strong differences between depleted participants and the control group. It would be desirable to measure ego-depletion over longer time frames, allowing to increase the strength of the manipulation. Of course, arguing that a manipulation was weak constitutes a general critique which could apply to all non-significant results in the ego-depletion literature. However, it is our impression that part of the interest in the ego depletion effects arose due to the impression that they would be induced by minimal, weak manipulations. If one were to conclude that ego depletion effects only appear for very strong manipulations, the theory would lose a large part of its appeal.

A second limitation is that, as most of the literature we relied on laboratory experiments. Even if the effects of ego depletion in the lab turn out to be too weak to be of general interest, we cannot exclude the possibility that they are relevant for naturally-occurring temptations. following up on this possibility requires a completely different approach. For instance, Hofmann, Baumeister, Förster, and Vohs (2012) registered naturally-occurring, everyday temptations of participants for a whole week.

A third limitation concerns the power necessary to detect ego depletion effects if those turn out to be very weak. To determine sample size, we started from the one used in the multi-lab registered replication of Hagger et al. (2016). Friese et al. (2018) argue that the effect size of $d = 0.62$ used there, as originally estimated by Hagger et al. (2010), might quite possibly be inflated due to publication bias. Our studies (with both $N > 200$) actually have a power of 0.82 for an small-to-moderate effect size ($d = 0.35$). Recent results by Garrison, Finley, and Schmeichel (2018) estimate an even smaller effect size of $d = 0.20$ for ego-depletion. Our studies lack the power necessary for identifying such a small effect (0.43). Hence, we cannot discard the possibility that ego depletion exists but its effects are very weak. Of course, if this is the case and the effects of ego depletion are so small, the relevance of the phenomenon and associated theory is greatly diminished.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

References

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2015). Money, Depletion, and Prosociality in the Dictator Game. *Journal of Neuroscience, Psychology, and Economics*, *8*(1), 1–14.

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2016). The Impact of Self-Control Depletion on Social Preferences in the Ultimatum Game. *Journal of Economic Psychology*, *53*, 1–16.

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2018). Social Preferences and Self-Control. *Journal of Behavioral and Experimental Economics*, *74*, 161–166.

Alós-Ferrer, C., Hügelschäfer, S., & Li, J. (2015). Self-Control Depletion and Decision Making. *Journal of Neuroscience, Psychology, and Economics*, *8*(4), 203–216.

Arber, M. M., Ireland, M. J., Feger, R., Marrington, J., Tehan, J., & Tehan, G. (2017). Ego depletion in Real-Time: An Examination of the Sequential-Task Paradigm. *Frontiers in Psychology*, *8*, 1672.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego Depletion: Is the Active Self a Limited Resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265.

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

Bertrams, A., & Dickäuser, O. (2009). Measurement of Dispositional Self-Control Capacity: A German Adaptation of the Short Form of the Self-Control Scale (SCS-KD). *Diagnostica*, *55*(1), 2–10.

Brown, D. M., & Bray, S. R. (2017). Graded Increases in Cognitive Control Exertion Reveal a Threshold Effect on Subsequent Physical Performance. *Sport, Exercise, and Performance Psychology*, *6*(4), 355–369.

Carter, E. C., & McCullough, M. E. (2013). Is Ego Depletion Too Incredible? Evidence for the Overestimation of the Depletion Effect. *Behavioral and Brain Sciences*, *36*, 683–684.

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*, 1–11.

Dawson, M. R. W. (1988). Fitting the ex-Gaussian Equation to Reaction Time Distributions. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 54–57.

Fennis, B. M., Janssen, L., & Vohs, K. D. (2008). Acts of Benevolence: A Limited-Resource Account of Compliance with Charitable Requests. *Journal of Consumer Research*, *35*(6), 906–924.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics*, *10*(2), 171-178.

Friese, M., Binder, J., Luechinger, R., Boesiger, P., & Rasch, B. (2013). Suppressing Emotions Impairs Subsequent Stroop Performance and Reduces Prefrontal Brain Activation. *PLOS ONE*, *8*(4), e60385.

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2018). Is Ego Depletion Real? An Analysis of Arguments. *Personality and Social Psychology Review*, *forthcoming*.

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*(1), 49–62.

Garrison, K. E., Finley, A. J., & Schmeichel, B. J. (2018). Ego Depletion Reduces Attention Control: Evidence From Two High-Powered Preregistered Experiments. *Personality and Social Psychology Bulletin*, *forthcoming*.

Gissubel, K., Beiramar, A., & Freire, T. (2018). The Ego Depletion Effect on Undergraduate University Students: A Systematic Review. *Motivation and Emotion*, *in press*.

Hagger, M. S., & Chatzisarantis, N. L. D. (2014). It is Premature to Regard the Ego-Depletion Effect as "Too Incredible". *Frontiers in Psychology*, *5*, 1-2.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego Depletion and the Strength Model of Self-Control: A Meta-Analysis. *Psychological Bulletin*, *136*(4), 495-525.

Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday Temptations: An Experience Sampling Study of Desire, Conflict, and Self-Control. *Journal of Personality and Social Psychology*, *102*(6), 1318–1335.

Inzlicht, M., & Gutsell, J. N. (2007). Running on Empty: Neural Signals for Self-Control Failure. *Psychological Science*, *18*(11), 933–937.

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

Karalunas, S. L., Huang-Pollock, C. L., & Nigg, J. T. (2013). Is Reaction Time Variability in ADHD Mainly at Low Frequencies? *Journal of Child Psychology and Psychiatry*, *54*(5), 536–544.

Muraven, M., Shmueli, D., & Burkley, E. (2006). Conserving Self-Control Strength. *Journal of Personality and Social Psychology*, *91*(3), 524–537.

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-Control as Limited Resource: Regulatory Depletion Patterns. *Journal of Personality and Social Psychology*, *74*(3), 774-789.

Richeson, J. A., & Trawalter, S. (2005). Why Do Interracial Interactions Impair Executive Function? A Resource Depletion Account. *Journal of Personality and Social Psychology*, *88*(6), 934–947.

Savani, K., & Job, V. (2017). Reverse Ego-Depletion: Acts of Self-Control Can Improve Subsequent Performance in Indian Cultural Contexts. *Journal of Personality and Social Psychology*, *113*(4), 589–607.

Schmeichel, B. J., & Vohs, K. D. (2009). Self-Affirmation and Self-Control: Affirming Core Values Counteracts Ego Depletion. *Journal of Personality and Social Psychology*, *96*(4), 770–782.

Sripada, C., Kessler, D., & Jonides, J. (2014). Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological Science*, *25*(6), 1227–1234.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of Personality*, *72*(2), 271-324.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2008). Making Choices Impairs Subsequent Self-Control: A Limited-Resource Account of Decision Making, Self-Regulation, and Active Initiative. *Journal of Personality and Social Psychology*, *94*(5), 883–898.