Strong Effort Manipulations Reduce Response Caution: A Preregistered Reinvention of the Ego-Depletion Paradigm

Accepted at Psychological Science

Hause Lin[1], Blair Saunders[2], Malte Friese[3], Nathan J. Evans[4], Michael Inzlicht[1,5]

[1]University of Toronto, Canada, Department of Psychology

[2]University of Dundee, United Kingdom, Department of Psychology

[3]Saarland University, Germany, Faculty of Empirical Humanities and Economics

[4]University of Amsterdam, The Netherlands, Department of Psychology

[5]Rotman School of Management, Canada

Correspondence concerning this manuscript should be addressed Hause Lin, Department of Psychology, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON, M1C 1A4, Canada

Contact: hause.lin@mail.utoronto.ca

Abstract

People feel tired or “depleted” after exerting mental effort. But preregistered studies often fail to observe effects of exerting effort on behavioral performance in the laboratory or elucidate the underlying psychology. We tested a new paradigm in four preregistered within-subjects studies (*N* = 686). An initial high (vs. low) demand task reliably elicited very strong effort phenomenology. Afterwards, participants completed a Stroop task. We used drift diffusion modeling to obtain the boundary (response caution) and drift rate (information processing speed) parameters. Bayesian analyses indicated that the high demand manipulation reduced boundary but not drift rate. Increased effort sensations further predicted reduced boundary. However, our demand manipulation did not affect subsequent inhibition, as assessed with traditional Stroop behavioral measures and additional diffusion model analyses for conflict tasks. Thus, effort exertion reduced response caution rather than inhibitory control, suggesting that after exerting effort, people disengage and become uninterested in exerting further effort.

Keywords: fatigue, ego depletion, self-control, drift diffusion model, Bayesian analysis

Strong Effort Manipulations Reduce Response Caution: A Preregistered Reinvention of the Ego-Depletion Paradigm

What are the consequences of exerting effort? Many people intuitively believe in ego depletion (Francis & Job, 2018), the idea that exerting effortful control depletes one's energy (Baumeister & Vohs, 2016). However, high-powered preregistered studies (Garrison, Finley, & Schmeichel, 2019; Hagger et al., 2016), meta-analyses (Carter, Kofler, Forster, & McCullough, 2015), and theoretical reviews (Friese, Loschelder, Gieseler, Frankenbach, & Inzlicht, 2019; Inzlicht & Friese, 2019) suggest laboratory depletion effects are small or potentially non-existent, and that previous work suffers from limitations such as ineffective experimental manipulations and low statistical power. Here, in four preregistered studies, we developed a paradigm that addresses previous methodological limitations and provides insights into the effects of effort exertion.

Ego Depletion Controversy

In the first tests of ego depletion (i.e., Baumeister, Bratslavsky, Muraven, & Tice, 1998), one group of participants initially completed a difficult self-control task (depletion group; e.g., forced to eat radishes instead of chocolates), while another group completed an easier task (control group; e.g., allowed to eat chocolates). Both groups then completed a second unrelated self-control task (e.g., worked on unsolvable puzzles), which served as the dependent variable (e.g., persistence duration on puzzles). The depletion (vs. control) group showed reduced self-control on the second task, providing evidence for ego depletion—the idea that self-control runs out after use. (Friese et al., 2019)

Subsequent studies found that depletion influenced diverse outcomes, even when the depletion or outcome tasks did not entail self-control or inhibitory control (e.g., Moller, Deci, & Ryan, 2006; Schmeichel, 2007), suggesting that exerting effort and experiencing fatigue (rather than recruiting inhibitory control) led to depletion effects. Critically, the first meta-analysis of 198 published tests suggested the effect (d^{+} = 0.62, 95% CI [0.57, 0.67]) was practically important and deserved further investigation (Hagger, Wood, Stiff, & Chatzisarantis, 2010).

Subsequent evidence, however, suggested otherwise. Published studies began reporting replication failures or much smaller effect sizes (e.g., Tuk, Zhang, & Sweldens, 2015). Meta-analyses that attempted to correct for publication bias (i.e., significant results are published more frequently than non-significant ones) suggest depletion might be unreal (Carter et al., 2015; Friese & Frankenbach, 2019), though subsequent work has questioned the validity of existing bias-correction techniques (Carter, Schönbrodt, Gervais, & Hilgard, 2019). These critiques were bolstered by further failures involving either large-scale preregistered replications or reanalyses of large datasets not originally gathered to investigate ego depletion (Etherton et al., 2018; Hagger et al., 2016).

Starting Anew: A Novel Approach

Although the field appears to have hit a dead end, it might be too soon to jettison ego depletion because the field has relied mainly on one paradigm (i.e., between-subjects laboratory sequential tasks) and has yet to fully examine other approaches. For example, studies using archival datasets, field data, or experience sampling suggest depletion or carry-over fatigue effects may be apparent in people's everyday lives (e.g., Dai, Milkman, Hofmann, & Staats, 2015; Hirshleifer, Levi, Lourie, & Teoh, 2019). Although ecologically valid, these studies often cannot control for real-world confounds. Our goal was to create a laboratory paradigm to provide converging evidence to facilitate future research. We also tested the idea that laboratory depletion effects are akin to real-life fatigue effects whereby people shift their priorities when tired, resulting in disengagement from ongoing tasks (Hockey, 2013; Inzlicht, Schmeichel, & Macrae, 2014).

Strong manipulation and within-subjects design. Instead of using standard depletion paradigms, which often use demanding tasks that are thought to tap inhibitory control, we focused on designing a manipulation that robustly elicited states (e.g., effort, fatigue) typically associated with depletion (Friese et al., 2019). We used the symbol-counting task, which draws on the shifting and updating aspects of executive function (Garavan, Ross, Li, & Stein, 2000).

Crucially, we modified the task, allowing it to adapt trial-by-trial to each participant's performance, ensuring the task was highly demanding for each participant. Second, we used a completely within-subjects design to reduce error variance and increase statistical power (Francis, Milyavskaya, Lin, & Inzlicht, 2018). To minimize demand characteristics and learning effects, participants completed the low and high demand tasks on two separate days, spaced roughly one week apart.

Drift diffusion modeling. After the demand manipulation, participants completed the Stroop task, which is often used to assess inhibition abilities (Miyake & Friedman, 2012). Importantly, in addition to performing traditional behavioral analyses on reaction time and accuracy, our primary interest was to transform these observed measures into latent variables assumed to underlie performance. We fitted drift diffusion models, which assume people make speeded decisions by gradually accumulating information until an evidence boundary is reached (Fig. 1; Ratcliff & McKoon, 2008). This model not only resolves the speed-accuracy trade-off in reaction-time tasks (Ratcliff & McKoon, 2008) but also allows us to examine if fatigue effects affect information processing speed (drift rate parameter) and response caution or impulsivity (boundary parameter; see Fig. 1 for explanation). Specifically, we fitted the EZ-diffusion model (Wagenmakers, van der Maas, & Grasman, 2007), which, despite being simpler, often outperforms the full diffusion model and better detects experimental effects (Dutilh et al., 2019; van Ravenzwaaij, Donkin, & Vandekerckhove, 2017). Given the success of this modeling approach in explaining individual differences and how experimental manipulations influence psychological processes (Evans & Wagenmakers, 2019), these latent variables can provide insights into the psychology underlying depletion.

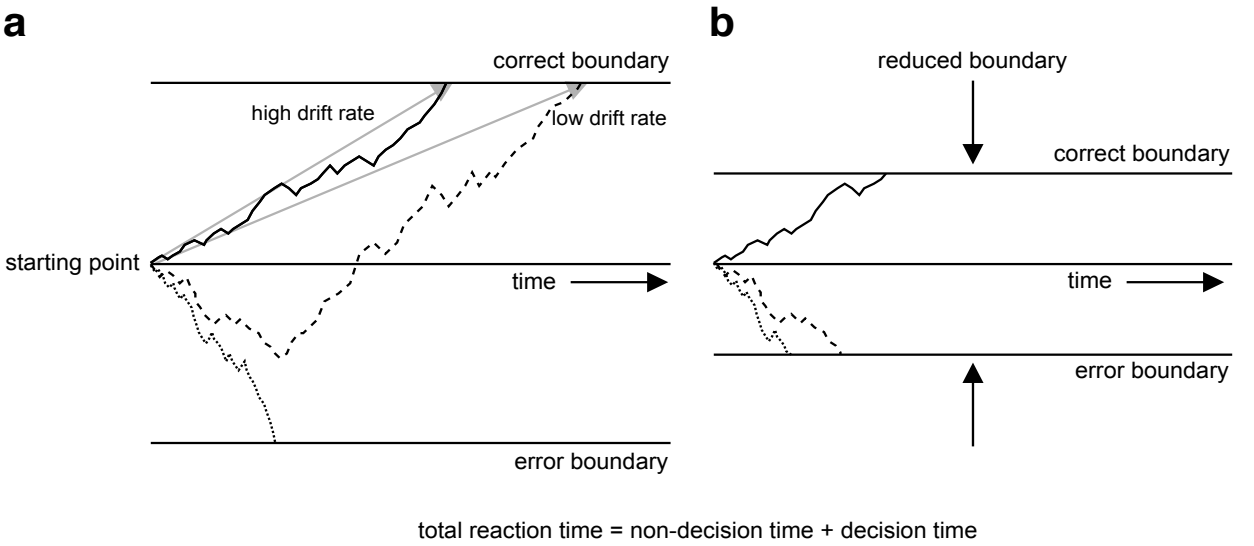

Fig. 1. The drift diffusion model decomposes the joint distributions of reaction time and accuracy into latent variables including drift rate, decision boundary, non-decision time, and starting point (Ratcliff & McKoon, 2008). Fig. 1a shows three simulated decisions with different diffusion processes or paths. Each path depicts how one decision process evolved over time. The solid and dashed black lines depict decision processes reaching the correct boundary (i.e., correct responses made) at different rates, with higher drift rates terminating sooner at the boundary (i.e., faster reaction times). The black dotted line depicts a process that terminated relatively quickly at the error boundary (i.e., fast error response). Fig. 1b illustrates what were to happen if the boundaries were reduced for the same three decisions. Decision processes terminate at the boundaries sooner even though drift rates remain unchanged, reflecting less evidence accumulation, resulting in noisier or more error responses and faster reaction times. The decision process depicted by the dashed line terminated prematurely at the error boundary. Boundary widths reflect either individual differences in response caution or experimental manipulations (e.g., emphasizing speedy responses reduces boundaries whereas emphasizing accuracy increases them).

Hypotheses

Studies 1 and 2 were conducted in the laboratory. Studies 3 and 4 were conducted online. We varied the high-demand task duration across studies. In Study 1, we preregistered only traditional analyses on the Stroop task but ran exploratory diffusion model analyses. Studies 2, 3, and 4 were preregistered, confirmatory experiments that tested two primary hypotheses: The high (vs. low) demand experimental manipulation would reduce boundary and drift rate. These predictions reflect our prediction that exerting effort should reduce subsequent overall task engagement, rather than specifically inhibitory control.

Method

Participants

Study 1 (N = 253 undergraduates; 178 females, 71 males, 4 others; age: M = 18.80, SD = 2.66; range = 17–46; https://osf.io/hhn3s/) was designed to primarily evaluate the effectiveness of our within-subjects demand manipulation and its effects on traditional Stroop behavioral measures (i.e., accuracy and reaction time). We also ran exploratory diffusion model analyses, which we planned to confirm and replicate in Studies 2 to 4, whereby we assumed a relatively small effect size (d = 0.26), which reflected our beliefs at the time of preregistration, skepticism around depletion research, and the likelihood that previous studies might have over-estimated effect sizes. We conducted sensitivity analyses using the Power Analysis for General Anova Designs R Shiny app (Westfall, 2015), which suggested that roughly 130 participants would provide at least 80% statistical power. We tried our best to recruit about 130 participants for each study, but since we recruited participants in batches and had to exclude data (see Exclusion Criteria section), our final sample sizes were not exactly 130: Study 2 (N = 132 undergraduates; 98 females, 32 males, 2 others; age: M = 18.80, SD = 1.78; range = 17–29; https://osf.io/xp7hn/); Study 3 (N = 180 MTurk workers; 94 females, 83 males, 3 others; age: M = 34.90, SD = 9.93; range = 20–70; https://osf.io/6p8t4/); Study 4 (N = 121 MTurk workers; 63 females, 57 males, 1 others; age: M = 39.50, SD = 11.20; range = 20–66; https://osf.io/6sncm/).

All participants provided informed consent in accordance with policies of the university's institutional review board.

Within-Subjects Design

To reduce error variance and increase statistical power, all four studies used within-subjects designs. To minimize demand characteristics and learning effects, each participant completed the low and high demand tasks on two separate days. In Studies 1 and 2, undergraduate participants completed the two tasks in two different weeks. Both sessions occurred on the same day of each week, at the same time of the day. They were pseudo-randomly assigned to complete either the low or high demand task on the first day based on their allocated participant number. Participants in Studies 1 and 2 received course credits for completing the study. In Studies 3 and 4, Amazon MTurk workers also completed the two tasks in two different weeks. However, since participants recruited via this online platform usually complete tasks at their convenience, they completed the second task 7 to 12 days after they completed the first task. They also did not have to complete the two tasks at same time of the day. They were randomly assigned to complete either the low or high demand task on the first day. Participants in Studies 3 and 4 received US$2.90 and $3.60 respectively for completing the study.

Procedure and Sequential-Task Paradigm

Task 1: Experimental manipulation. The low demand task required participants to watch a 5-minute wildlife video. The high demand task required participants to complete a titrated symbol-counting task (study materials and code available here: git.io/JeDaH; see Garavan et al., 2000) that lasted approximately 20, 15, 5, and 10 minutes in Studies 1, 2, 3, and 4, respectively. We did not match the durations of the low and high demand tasks in Studies 1, 2, and 4 because we wanted to avoid inducing boredom with long but easy control tasks, which

might lead to comparable levels of subjective fatigue as exerting cognitive effort on a demanding task (Milyavskaya, Inzlicht, Johnson, & Larson, 2019), potentially undermining the demand manipulation. Further, previous work using unbalanced designs like ours have reported stronger effects (Sjåstad & Baumeister, 2018).

The symbol-counting task is a cognitive task that parametrically manipulates executive demands (Garavan et al., 2000). On each trial, participants had to count the number of small black squares that had been presented. As such, the task heavily taxes the shifting and updating (but not inhibition) aspects of executive function (Miyake & Friedman, 2012). To further increase the difficulty of the task, we calibrated the task for each individual such that it adjusted its difficulty trial-by-trial according to the participant's performance on the previous trial. On each trial, multiple small and big squares were presented sequentially (between 11 to 17 squares per trial), and each square was preceded by a fixation cross (Fig. 2). The first trial began with 12 squares and a switch frequency of 5 (i.e., the squares within a trial switched 5 times, from small to big or big to small square). After all squares were presented, participants indicated how many small and black squares were presented. That is, participants had to keep a running tally of two lists. If participants responded correctly, the total number of squares in the next trial increased by one, the switch could also increase, and the square display duration decreased by 20 ms (see Table S1 in the Supplemental Online Material Reviewed for details on how the switch frequency was determined on each trial and other task details). If participants responded incorrectly, the number of squares on the next trial decreased by one, the square display duration increased by 20 ms, and the switched frequency decreased. These calibration procedures helped to ensure that even without drawing on inhibition processes, the task was demanding and tiring for all participants, regardless of individual differences in executive function abilities.

Measures of Phenomenology. After completing the low and high demand tasks, participants answered five questions (presented in random order) about the task they had just

completed and their current mental state using a sliding Likert scale (two anchors provided in parentheses): (1) mental demand: “How mentally demanding was the (video) task?” (very low demand, very demanding); (2) effort: "How hard did you have to work (to watch the video)?” (very little, very hard); (3) frustration: “How insecure, discouraged, irritated, stressed, and annoyed were you (when watching the video)?” (very little, very high); (4) boredom: “How boring was the (video) task?” (not boring, very boring); (5) fatigue: “I’m mentally fatigued now.” (strongly disagree, strongly agree). Each scale ranged from 1 to 7, but participants did not see the scale ranges and only saw the two text anchors (e.g., not boring, very boring) below the sliding Likert scale.

Task 2: Outcome measure. After completing the experimental manipulation and manipulation checks, participants completed a Stroop task with 120 congruent and 60 incongruent trials. On each trial, a word (“red”, “blue”, or “yellow”) was presented in either the color red, blue, or yellow, and participants had to indicate the font color of the word by making a keypress (V: red; B: blue; N: yellow). The same mapping was used for all participants and was displayed at the bottom of the screen throughout the task. On congruent trials, the word and color matched (the word “red” shown in red font); on incongruent trials, the word and color did not match (the word “red” shown in blue font). Congruent and incongruent trials were interleaved randomly and the stimulus on each trial remained on screen until the participant responded or until a maximum of 2000 ms. If participants failed to respond on three consecutive trials, they were reminded to respond faster and more accurately. Participants practiced 12 trials before completing 180 experimental trials.

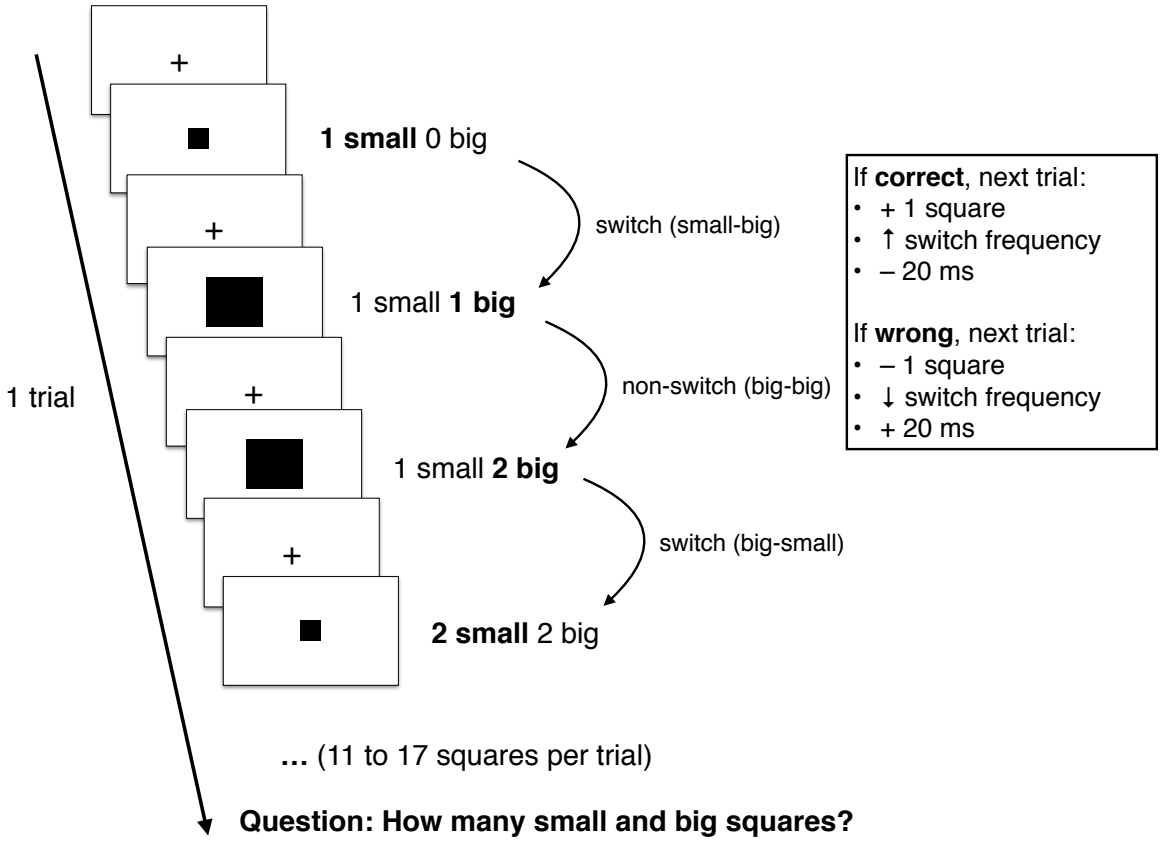

Fig. 2. Titrated symbol counter task as the high demand manipulation (adapted from Garavan et al., 2000). This calibrated task heavily taxes the shifting and updating aspects of executive function. On each trial, multiple small and big squares were presented sequentially, and participants reported the number of small and big squares presented at the end of the trial. If participants responded correctly, the total number of squares in the next trial increased, the switch frequency increased, and the square display duration decreased. If participants responded incorrectly, the total number of squares on the next trial decreased, the switch frequency decreased, and the square display duration increased.

Exclusion Criteria

We used the same four preregistered exclusion criteria[1] for all four studies to exclude low quality data (e.g., see Exclusion criteria section at https://osf.io/6sncm/). First, we excluded participants whose overall accuracy on the high demand task (titrated symbol counter task) was less than 20% (3, 9, 25, 28 participants excluded in Studies 1, 2, 3, 4 respectively). Second, for the dependent variable (Stroop task), we excluded trials whereby reaction time was faster than 250 ms (0.52%, 0.26%, 1.66%, 0.93% trials excluded in Studies 1, 2, 3, 4 respectively). Third, we used a robust outlier detection approach (median absolute deviation) rather than the commonly-used but problematic ±3 SD approach to exclude trials with outlier reaction times (Leys, Delacre, Mora, Lakens, & Ley, 2019). For each participant, within each experimental condition, we excluded trials whereby the reaction time was the median ±3 times the median absolute deviation (5.44%, 5.06%, 5.29%, 4.68% trials excluded in Studies 1, 2, 3, 4 respectively). Fourth, we used the same robust approach and criterion to exclude participants who made too many errors on congruent Stroop trials (12, 3, 24, 16 participants excluded in Studies 1, 2, 3, 4 respectively). Note that these four criteria did not pertain to our dependent variables because the goal was to exclude extremely low-quality data (e.g., disengaged or inattentive participants who showed little signs of trying), rather than to exclude outliers based on the outcome variables. Rerunning our main analyses with outliers re-included did not change our main conclusions (Supplemental Online Material Unreviewed Table S4).

Diffusion Model Fitting

We fitted the EZ-diffusion model (Wagenmakers et al., 2007) to each participant's Stroop data, which transformed the observed reaction time and accuracy variables into the latent

[1] We preregistered additional criteria in exploratory Study 1. However, in confirmatory Studies 2 to 4, we focused on the latent variables and did not preregister criteria based on behavioral effects. For consistency across studies, we applied to Study 1 the criteria used in Studies 2 to 4.

variables of drift rate, boundary, and non-decision time (for code, see Lin, 2019). This model does not compute the starting point bias because it assumes the starting point is equidistant from the two boundaries. Furthermore, the boundary parameter is generally assumed to be determined prior to stimulus onset and therefore should not vary as a function of Stroop stimulus congruency. However, using the EZ-diffusion model in our case prevents us from forcing the boundary to be the same for all stimuli. We therefore obtained separate boundary parameter estimates for congruent and incongruent Stroop stimuli, assuming that participants rapidly adjust their boundaries immediately after stimulus onset.

Despite these assumptions and being a simpler model, several studies have shown that the EZ-diffusion model often outperforms the full diffusion model (Dutilh et al., 2019; van Ravenzwaaij et al., 2017). To verify the EZ-diffusion model results, we ran exploratory but preregistered analyses (osf.io/7qcxa) to fit more appropriate models (i.e., diffusion model for conflict tasks; Evans & Servant, 2019), which led to conclusions similar to those obtained via EZ-diffusion modeling (see Supplemental Online Material Unreviewed Figures S2 and S3).

Preregistered Hypotheses and Analyses

Phenomenology. We expected participants to report higher mental demand, effort exerted, frustration, boredom, and fatigue in the high than low demand condition.

Primary hypotheses. We expected the high demand condition to have a smaller boundary than the low demand condition, after controlling for Stroop congruency (trial type: congruent or incongruent), reflecting less cautious or more impulsive responding after completing the high demand task. We also expected the high demand condition to have lower drift rate than the low demand condition, after controlling for Stroop congruency, reflecting slower information processing rate. These analyses reflect our belief at the time that our demand manipulation should reduce overall task motivation and engagement, rather than reduce specifically self-control or inhibition abilities. Note that we preregistered these two

hypotheses only in Studies 2 to 4 but not Study 1, where we only preregistered traditional Stroop behavioral effects. Finally, another plausible outcome[2] we did not preregister (and failed to observe in our data) is that effort exertion reduces drift rate, and participants might compensate by increasing boundary separation to ensure they maintain acceptable accuracies on the task.

Secondary hypotheses. We also tested additional hypotheses that indirectly examined the effects of high versus low demand, but the primary effects described above did not hinge on these secondary effects. Based on the results from Study 1, we expected that (1) participants who reported feeling more fatigued, frustrated, or bored[3] after completing the first task would have lower drift rate or boundary on the Stroop task; and (2) incongruent Stroop trials would be associated with lower drift rate and boundary[4] than congruent Stroop trials.

Exploratory Analyses

[2] We thank Eric-Jan Wagenmakers for highlighting this hypothesis, which increases the informativeness of our results by highlighting which processes were (not) influenced by effort exertion.

[3] When exploring Study 1's data, we found hints of associations between the two latent parameters and three items assessing phenomenology (frustration, bored, fatigue) but not two others (mental demand, effort exerted). We thus pre-registered analyses with frustration, bored, and fatigues, but consider the analyses with mental demand and effort exploratory.

[4] We made this prediction based on Study 1's results. Theoretically, we would have made the opposite prediction because incongruent trials are more difficult, and boundaries should increase to allow more time for evidence accumulation. Reduced boundaries on incongruent trials could reflect reduced drift rates on the same trials: When drift rates are low, evidence accumulation rates are slower, potentially increasing in missed responses, which could be avoided by lowering the boundaries so the decision processes can still reach a bound within the allotted time.

We investigated the effects of our manipulations on traditional Stroop behavioral outcomes (reported in the main text). Note that we preregistered these behavioral effects in Study 1, but not Studies 2 to 4. In addition, we verified the EZ-diffusion model results by running exploratory but preregistered analyses (osf.io/7qcxa) involving fitting more complicated diffusion models (Evans & Servant, 2019; Ulrich, Schröter, Leuthold, & Birngruber, 2015). Finally, because Stroop performance might be influenced by practice or learning effects (due to our within-subjects design), we also tested for session-order effects.

Statistical Analyses

Continuous predictors were participant mean-centered and categorical predictors were recoded prior to model fitting: condition (low demand: −0.5; high demand: 0.5) and Stroop congruency (congruent: −0.5; incongruent: 0.5). We fitted Bayesian multilevel models using the R package brms (Bürkner, 2017). We first fitted two-level varying-intercept multilevel models separately for each study whereby data/units clustered within participants ($y_i = \beta_{0\,[participant][i]} + X_i\beta + \epsilon_{[i]}$; R syntax: *(1 | participant)*). To meta-analyze the four studies to obtain an overall effect, we fitted three-level varying-intercept multilevel models whereby data were clustered within participants, who were in turn clustered within studies ($y_i = \beta_{0\,[study][participant][i]} + X_i\beta + \epsilon_{[i]}$; R syntax: *(1 | study/participant)*).

For the condition effect (high vs. low demand) in each model, we used an informed Gaussian prior of $d = 0.28$ ($SD = 0.14$), which was based on a Bayesian reanalysis of a published depletion study (Wagenmakers & Gronau, 2017). The priors were rescaled to the raw scale of each outcome measure, such that the prior mean reflected the expected difference between low and high demand conditions, and the standard deviation of the prior distribution was half the prior mean (Dienes, 2014). For example, for the effects of condition on self-reported demand and boundary, the priors were $N(0.36, 0.18)$ and $N(-0.0088, 0.0044)$ respectively (see Fig. 4 in main text and Fig. S1 in the Supplemental Online Material

Unreviewed for visualizations of the prior and posterior distributions). For other effects that did not directly test the effect of our demand manipulation, we used the standard normal prior $N(0, 1)$.

Since the prior influences the posterior, we performed prior sensitivity analyses by refitting the models using normal priors with the same standard deviations as the informed priors but centered around 0 for the effect of interest; effects not directly testing our demand manipulation had the prior $N(0, 1)$. These priors reflected the belief that our experimental effects would be relatively tightly centered around 0: For example, the priors for the effects of condition on self-reported mental demand and boundary were $N(0, 0.18)$ and $N(0, 0.0044)$ respectively (compare with informed priors above). Results from the sensitivity analyses were consistent with our main or original conclusions, suggesting our findings were robust to prior choice (see Table S2 in the Supplemental Online Material Reviewed for complete results from models fitted using these priors).

For each model, we ran 20 MCMC chains with 2000 samples and discarded the first 1000 samples (burn-in). For each effect, we reported the mean of the posterior samples and the 95% highest posterior density interval (HPD), which is the narrowest interval containing the specified probability mass. We used bridge sampling to compute Bayes factors (BF), which reflects the amount of evidence favoring one model over a reduced model that did not contain the effect or hypothesis of interest. To ensure the stability of the results, the reported BFs were the mean of five BF computations. Here, BF = 1 indicates equal evidence favoring the null and experimental hypotheses and BF > 1 indicate evidence in favor of the experimental hypothesis: 1 to 3 indicates anecdotal evidence, 3 to 10 indicates moderate evidence, 10 to 30 indicates strong evidence, and > 30 indicates very strong or decisive evidence (Lee & Wagenmakers, 2013; but for problems with Bayes factors see). Conversely, BF < 1 indicates evidence in favor of the null hypothesis, with smaller values indicating stronger evidence for the null hypothesis:

0.33 to 1 indicates anecdotal evidence, 0.10 to 0.33 indicates moderate evidence, 0.03 to 0.10 indicates strong evidence, and BF < 0.03 indicates very strong or decisive evidence.

All data, materials, and code for the main analyses can be found here: https://github.com/hauselin/depletion_bayes

Preregistered Analysis Results

Phenomenology

Demand. We found strong and consistent effects of condition on self-reported mental demand. In all studies (Fig. 3), mental demand was much higher in the high than low demand condition (Study 1: b = 1.96, 95% HPD [1.75, 2.17], d = 1.46; Study 2: b = 1.73, 95% HPD [1.45, 2.01], d = 1.21; Study 3: b = 2.23, 95% HPD [1.96, 2.48], d = 1.56; Study 4: b = 2.06, 95% HPD [1.73, 2.39], d = 1.37; see Table 1 for more information), suggesting that our paradigm was highly effective for eliciting effort-related phenomenology. To meta-analyze the effects across studies, we fitted a three-level multilevel model (data/units clustered within participants, who were clustered within studies). The meta-analytic effect was equally strong, b = 2.83, 95% HPD [2.70, 2.96], BF > 500, d = 2.17, 95% HPD [2.02, 2.33] (see Fig. 3). Self-reported demand was much higher in the high demand condition (M_{high} = 5.61, SD_{high} = 1.23) than the low demand condition (M_{low} = 2.41, SD_{low} = 1.36).

Effort. Self-reported effort was also much higher in the high than low demand condition in all studies (Study 1: b = 1.90, 95% HPD [1.69, 2.11], d = 1.42; Study 2: b = 1.58, 95% HPD [1.31, 1.86], d = 1.11; Study 3: b = 2.37, 95% HPD [2.11, 2.63], d = 1.72; Study 4: b = 1.99, 95% HPD [1.66, 2.31], d = 1.29). Results from the three-level multilevel model meta-analysis suggest that, overall, participants reported exerting much more effort in the high than low demand condition, b = 2.75, 95% HPD [2.61, 2.88], BF > 500, d = 2.09, 95% HPD [1.94, 2.24] (M_{low} = 2.33, SD_{low} = 1.42; M_{high} = 5.44, SD_{high} = 1.22).

Frustration. Similarly, participants reported feeling more frustrated in the high than low demand condition in all studies (Study 1: b = 2.13, 95% HPD [1.91, 2.36], d = 1.60; Study 2: b = 1.31, 95% HPD [1.02, 1.59], d = 0.88; Study 3: b = 1.50, 95% HPD [1.23, 1.77], d = 0.97; Study 4: b = 1.48, 95% HPD [1.15, 1.78], d = 0.96). The three-level multilevel model meta-analysis results were similar, b = 2.10, 95% HPD [1.95, 2.24], BF > 500, d = 1.50, 95% HPD [1.36, 1.64] (M_{low} = 2.19, SD_{low} = 1.41; M_{high} = 4.49, SD_{high} = 1.70).

Boredom. Self-reported boredom was higher in the high than low demand condition in all studies (Study 1: b = 1.13, 95% HPD [0.89, 1.38], d = 0.72; Study 2: b = 0.69, 95% HPD [0.38, 1.00], d = 0.43; Study 3: b = 0.60, 95% HPD [0.31, 0.89], d = 0.35; Study 4: b = 0.74, 95% HPD [0.40, 1.08], d = 0.41). Results from the three-level multilevel model meta-analysis suggest the effect was consistent across studies but smaller than the effects on demand, effort, and frustration, b = 0.91, 95% HPD [0.75, 1.08], BF > 500, d = 0.55, 95% HPD [0.45, 0.66] (M_{low} = 3.53, SD_{low} = 1.83; M_{high} = 4.48, SD_{high} = 1.87).

Fatigue. Finally, participants reported higher fatigue in the high than low demand condition in all studies (Study 1: b = 2.05, 95% HPD [1.84, 2.25], d = 1.65; Study 2: b = 1.41, 95% HPD [1.13, 1.69], d = 0.97; Study 3: b = 1.92, 95% HPD [1.65, 2.19], d = 1.25; Study 4: b = 1.98, 95% HPD [1.64, 2.31], d = 1.27). Results from the three-level multilevel model meta-analysis were similar, b = 2.49, 95% HPD [2.35, 2.63], BF > 500, d = 1.86, 95% HPD [1.70, 2.02], suggesting that our high demand task was effective in eliciting fatigue (M_{low} = 2.11, SD_{low} = 1.30; M_{high} = 4.87, SD_{high} = 1.58).

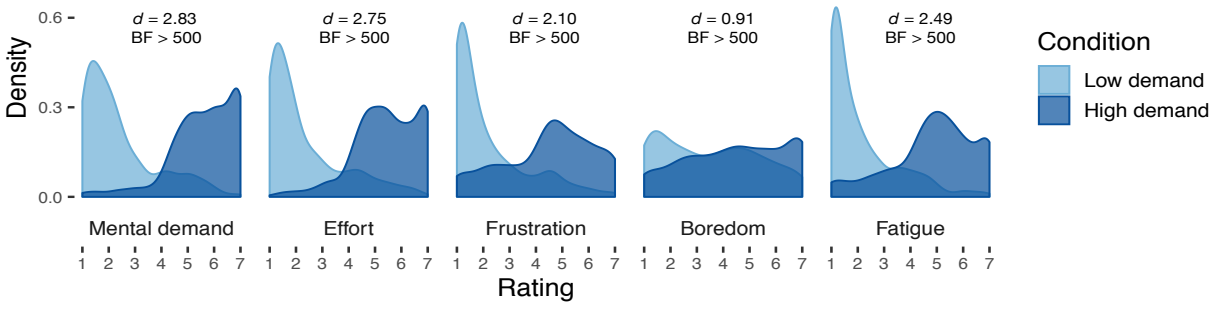

Figure 3. Phenomenology collapsed across studies. Self-reported ratings of mental demand, effort, frustration, boredom, and fatigue were much higher in the high than low demand condition. Kernel density estimates are shown (area beneath each density curve sum to 1). See Table 1 for detailed statistics.

Latent Parameter: Boundary

We report the effect of condition on the boundary parameter after controlling for Stroop congruency, which had strong effects on the boundary parameter in all studies (Fig. 4, Table 1). The boundary parameter was smaller in the high (vs.) low demand condition in all four studies (Study 1: b = −0.004, 95% HPD [−0.006, −0.002], d = −0.22; Study 2: b = −0.002, 95% HPD [−0.005, 0.001], d = −0.08; Study 3: b = −0.005, 95% HPD [−0.008, −0.003], d = −0.32; Study 4: b = −0.006, 95% HPD [-0.01, 0.00], d = −0.13), but only Studies 1 and 3 had effects whose 95% HPD did not include 0. Moreover, results from the three-level multilevel model meta-analysis provided strong evidence for our preregistered hypothesis that completing a high (vs. low) demand task would lead to reduced boundary, b = −0.004, 95% HPD [−0.005, −0.002], BF = 29.10, d = −0.15, 95% HPD [−0.22, −0.07]; this effect was similar when we used a prior centered around 0 (but retaining the scale of the informed prior), b = −0.003, 95% HPD [−0.005, −0.001], BF = 88.33, d = −0.13, 95% HPD [−0.20, −0.06]. Together, our results provide strong and decisive evidence in favor of the hypothesis that exerting mental effort decreases subsequent boundary separation. Nonetheless, even if reliable, the meta-analytic effect size was small (b = −0.004, d = −0.13) and slightly less than half the expected effect size (prior b = −0.0088, prior d = −0.28; see Fig. 4).

Exploratory analyses including session order and the order-condition interaction in the models showed that practice or learning effects were strong and consistent with previous work (e.g., Dutilh, Krypotos, & Wagenmakers, 2011). Boundary separation was smaller in the second than first session in all studies (BFs > 500), but order did not interact with condition, and the

effect of our demand manipulation remained small, but highly robust, b = −0.003, 95% HPD [−0.005, −0.002], BF = 117.67, d = −0.14, 95% HPD [−0.21, −0.07] (see Supplemental Online Material Unreviewed Figure S6 and Table S1).

Latent Parameter: Drift Rate

We also report the effect of condition on drift rate after controlling for Stroop congruency, which had strong effects on the drift rate parameter in all studies (Fig. 4, Table 1). The effects of condition on drift rate were inconsistent across studies (Study 1: b = −0.007, 95% HPD [−0.01, −0.001], d = −0.14; Study 2: b = −0.01, 95% HPD [−0.02, −0.002], d = −0.20; Study 3: b = 0.001, 95% HPD [−0.006, 0.009], d = 0.03; Study 4: b = −0.003, 95% HPD [−0.01, 0.007], d = −0.05). Further, results from the three-level multilevel model meta-analysis suggest—contrary to our preregistered hypothesis—that completing a high (vs. low) demand task did not lead to reduced drift rate, b = −0.003, 95% HPD [−0.007, 0.001], BF = 0.06, d = −0.05, 95% HPD [−0.13, 0.02] (Fig. 4).

Exploratory analyses including session order and order-condition interaction in the models showed that practice or learning effects were strong and these effects were consistent with previous work (e.g., Dutilh et al., 2011). Drift rate was higher in the second than first session in all studies (BFs > 500), reflecting improved task performance. Order did not interact with condition, and drift rate did not differ between the high and low demand conditions (see Supplemental Online Material Unreviewed Figure S6 and Table S1).

Finally, to verify the EZ-diffusion model results, we ran exploratory analyses that fitted the diffusion model for conflict tasks, which is specifically designed for cognitive control tasks like the Stroop. It models information integration during conflict tasks as a function of controlled and automatic processes, with the automatic process varying over time according to a gamma function (Ulrich et al., 2015). Consistent with the EZ-diffusion model results, effort exertion reduced boundary separation but had no effect on information integration via controlled (drift

rate parameter) or automatic processes (ζ parameter). These results bolster our interpretation that exerting effort or being "depleted" does not selectively impair one's ability to inhibit automatic processes (see Supplemental Online Material Unreviewed Figures S2 to S5 for details and additional results from the regular analytic diffusion model).

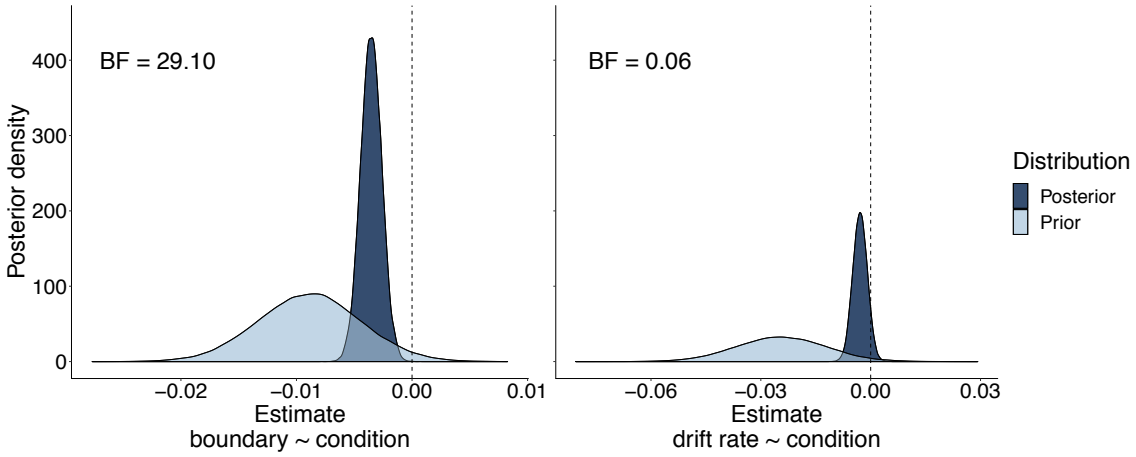

Figure 4. Bayesian posterior and prior density distributions for the effect of condition (low vs. high demand) on the boundary (left) and drift rate (right) parameters obtained from meta-analytic Bayesian multi-level models. Prior distributions reflect beliefs of the effect sizes prior to collecting empirical data: Informed priors reflecting Cohen's *d* = −0.28 (*SD* = 0.14) were created by rescaling the expected effect size to the raw scale of each parameter. Posterior distributions reflect the revised or updated beliefs and effect sizes after taking into consideration empirical data. Bayes factors (BF) suggest strong evidence in favor of the experimental hypothesis for the effect of condition on boundary (left) but more evidence in favor of the null hypothesis for the effect of condition on drift rate (right). See Table 1 for detailed statistics.

Phenomenology and Latent Parameter Relations

Boundary-fatigue relation. Increased self-reported fatigue was associated with smaller boundaries in three studies (Study 1: *b* = −0.001, 95% HPD [−0.002, 0.00], *d* = −0.09; Study 3: *b* = −0.002, 95% HPD [−0.003, −0.001], *d* = −0.14; Study 4: *b* = −0.002, 95% HPD [−0.004,

0.00], d = −0.04) but not Study 2 (Study 2: b = 0.00, 95% HPD [−0.001, 0.001], d = 0.01). Further, results from the three-level multilevel model meta-analysis provided weak evidence for the prediction that increased fatigue was associated with reduced boundary, b = −0.001, 95% HPD [−0.002, 0.00], BF = 1.88, d = −0.05, 95% HPD [−0.08, −0.02].

Boundary-frustration relation. Increased self-reported frustration was associated with smaller boundaries in three studies, although only Study 3's 95% HPD intervals did not include 0 (Study 1: b = −0.001, 95% HPD [−0.002, 0.00], d = −0.08; Study 2: b = 0.00, 95% HPD [−0.001, 0.002], d = 0.02; Study 3: b = −0.002, 95% HPD [−0.003, −0.001], d = −0.12; Study 4: b = −0.002, 95% HPD [−0.005, 0.001], d = −0.04). Overall, the three-level multilevel model meta-analysis results indicated that frustration was not associated with reduced boundary, b = −0.001, 95% HPD [−0.002, 0.00], BF = 0.25, d = −0.04, 95% HPD [−0.08, −0.009].

Boundary-boredom relation. Self-reported boredom was not associated with the boundary parameter in all four studies. All 95% HPD intervals included 0 (Table 1).

Drift rate-fatigue relation. Self-reported fatigue was not associated with the drift rate parameter in all four studies. All 95% HPD intervals included 0 (Table 1).

Drift rate-frustration relation. Self-reported frustration was not associated with the drift rate parameter in all four studies. All 95% HPD intervals included 0 (see Table 1).

Drift rate-boredom relation. Self-reported boredom was associated with reduced drift rate in Studies 1 and 2 (Study 1: b = −0.004, 95% HPD [−0.006, −0.001], d = −0.10; Study 2: b = −0.003, 95% HPD [−0.007, −0.001], d = −0.08; Study 3: b = −0.001, 95% HPD [−0.005, 0.002], d = −0.03; Study 4: b = −0.005, 95% HPD [−0.01, 0.00], d = −0.09). However, overall, the three-level multilevel model results failed to provide evidence for the prediction that increased boredom was associated with reduced drift rate, b = −0.003, 95% HPD [−0.004, −0.001], BF = 0.67, d = −0.06, 95% HPD [−0.11, −0.02].

Exploratory analyses found that session order did not interact with fatigue, frustration, or boredom (see Supplemental Online Material Unreviewed Tables S2 and S3).

Exploratory Analysis Results

Other (non-preregistered) analyses might provide further insights into the psychology of effort exertion and ego depletion (see Table 2). Note that because we did not preregister the analyses below, we used normal priors centered around 0 instead of informed priors.

Phenomenology and Latent Parameter Relations

Boundary-demand relation. Results from the individual studies showed that self-reported demand was not consistently associated with changes in boundary (Study 1: b = −0.001, 95% HPD [−0.002, 0.00], d = −0.05; Study 2: b = 0, 95% HPD [−0.001, 0.00], d = −0.03; Study 3: b = −0.001, 95% HPD [−0.002, 0.00], d = −0.10; Study 4: b = −0.001, 95% HPD [−0.003, 0.001], d = −0.03). Critically, the three-level multilevel model meta-analysis results provided strong evidence for a negative relationship between demand and boundary, b = −0.001, 95% HPD [−0.002, −0.001], BF = 17.21, d = −0.05, 95% HPD [−0.08, −0.02], though the effect was very small.

Boundary-effort relation. As with the boundary-demand results in the previous section, results from the individual studies for the boundary-effort relationship were mixed (Study 1: b = 0, 95% HPD [−0.001, 0.00], d = −0.04; Study 2: b = 0, 95% HPD [−0.002, 0.00], d = −0.03; Study 3: b = −0.001, 95% HPD [−0.002, −0.00], d = −0.10; Study 4: b = −0.001, 95% HPD [−0.004, 0.00], d = −0.03). However, the three-level multilevel model meta-analysis results also provided evidence for a negative relationship between self-reported effort and boundary, b = −0.001, 95% HPD [−0.002, 0.00], BF = 16.04, d = −0.05, 95% HPD [−0.08, −0.02], though the effect was also very small.

These two results above suggest that increased feelings of mental demand and effort led to less cautious responding on the Stroop task, though we note that these two ratings correlated strongly in all four studies (rs > .77). Since we consider these analyses exploratory, we caution

against over-interpreting these effects and have presented them because we believe they could provide insights into the effects of different subjective mental states.

Drift rate-demand relation. Self-reported demand was not associated with drift rate in all studies. All 95% HPD intervals included 0 (Table 2).

Drift rate-effort relation. Self-reported effort was also not associated with drift rate in all studies. All 95% HPD intervals included 0 (Table 2).

Latent Parameter: Boundary (Condition-Congruency Interaction)

We fitted a model whereby the boundary parameter was predicted by condition, Stroop congruency, and their interaction. Here, we focused on the interaction term because it indicates whether the effect of condition on the boundary parameter varied as a function of Stroop congruency. In all four studies, the 95% HPD intervals of the interaction term included 0 (see Table 2). Results from the three-level multilevel model were similar, $b = 0$, 95% HPD [−0.003, 0.004], BF = 0.43, $d = 0.03$, 95% HPD [−0.11, 0.17]. These findings suggest that the effect of condition on the boundary parameter did not vary as a function of Stroop congruency.

Latent Parameter: Drift Rate (Condition-Congruency Interaction)

We also fitted a model whereby the drift rate parameter was predicted by condition, Stroop congruency, and their interaction. In all studies, the interaction effect was close to 0 (Table 2), suggesting that the effect of condition on the drift rate parameter did not vary as a function of Stroop congruency.

Behavioral Effects: Stroop Accuracy

We modelled Stroop accuracy (proportion correct) as a function of condition, congruency, and their interaction. The congruency effect was robust and consistent across studies (Table 2) and the three-level multilevel model results indicate that accuracy was lower on incongruent trials, $b = -0.08$, 95% HPD [−0.08, −0.07], BF > 500, $d = -1.16$, 95% HPD [−1.24, −1.08] ($M_{congruent} = 0.98$, $SD_{congruent} = 0.03$; $M_{incongruent} = 0.90$, $SD_{incongruent} = 0.11$). The

condition effect (high vs. low demand) was negative in all studies, but all 95% HPD intervals included 0; however, results from the three-level multilevel model meta-analysis suggest there was some evidence for reduced overall accuracy in the high demand condition, though the effect was small, b = −0.005, 95% HPD [−0.01, 0.00], BF = 1.21, d = −0.08, 95% HPD [−0.15, −0.01] (M_{low} = 0.941, SD_{low} = 0.09; M_{high} = 0.935, SD_{high} = 0.09). The congruency-condition interaction effect was close to 0 in all studies and all 95% HPD intervals included 0 (Table 2), suggesting that the condition effect did not vary as a function of Stroop congruency.

Behavioral Effects: Stroop Reaction Time

We also modelled Stroop reaction time (correct trials) as a function of congruency, condition, and their interaction. As expected, the congruency effect was strong and consistent across studies (Table 2): Results from the three-level multilevel model meta-analysis indicate that reaction times on incongruent trials were slower on incongruent (vs. congruent) trials, b = 0.11, 95% HPD [0.10, 0.11], BF > 500, d = 1.73, 95% HPD [1.63, 1.82] ($M_{congruent}$ = 0.62, $SD_{congruent}$ = 0.10; $M_{incongruent}$ = 0.72, $SD_{incongruent}$ = 0.14). However, evidence for the condition effect (high vs. low demand) was less strong and mixed: The 95% HPD intervals for Studies 3 and 4 did not include 0 (Study 3: b = −0.01, 95% HPD [−0.02, −0.003], d = −0.19; Study 4: b = −0.01, 95% HPD [−0.02, −0.004], d = −0.25), whereas those of Studies 1 and 2 included 0 (see Table 2). The three-level multilevel model meta-analysis results provided some evidence that across the four studies, overall reaction times were faster in the high (vs. low) demand condition, though the effect was small, b = −0.006, 95% HPD [−0.01, −0.001], BF = 2.32, d = −0.09, 95% HPD [−0.17, −0.02] (M_{low} = 0.14, SD_{low} = 0.09; M_{high} = 0.66, SD_{high} = 0.13). The congruency-condition interaction effect was close to 0 in all studies (all 95% HPD intervals contained 0; see Table 2).

Table 1
Preregistered Analyses: Bayesian Multilevel Model Parameter Estimates Using Informed Priors

	Study 1 (20 min)	Study 2 (15 min)	Study 3 (5 min)	Study 4 (10 min)	Overall
Demand ~ condition	1.96 [1.75, 2.17] (BF > 500) *d* = 1.46 [1.25, 1.69]	1.73 [1.45, 2.01] (BF > 500) *d* = 1.21 [0.93, 1.49]	2.23 [1.96, 2.48] (BF > 500) *d* = 1.56 [1.29, 1.84]	2.06 [1.73, 2.39] (BF > 500) *d* = 1.37 [1.03, 1.72]	2.83 [2.70, 2.96] (BF > 500) *d* = 2.17 [2.02, 2.33]
Effort ~ condition	1.90 [1.69, 2.11] (BF > 500) *d* = 1.42 [1.20, 1.64]	1.58 [1.31, 1.86] (BF > 500) *d* = 1.11 [0.85, 1.38]	2.37 [2.11, 2.63] (BF > 500) *d* = 1.72 [1.42, 2.01]	1.99 [1.66, 2.31] (BF > 500) *d* = 1.29 [0.98, 1.62]	2.75 [2.61, 2.88] (BF > 500) *d* = 2.09 [1.94, 2.24]
Frustration ~ condition	2.13 [1.91, 2.36] (BF > 500) *d* = 1.60 [1.34, 1.87]	1.31 [1.02, 1.59] (BF > 500) *d* = 0.88 [0.65, 1.13]	1.50 [1.23, 1.77] (BF > 500) *d* = 0.97 [0.75, 1.22]	1.48 [1.15, 1.78] (BF > 500) *d* = 0.96 [0.68, 1.26]	2.10 [1.95, 2.24] (BF > 500) *d* = 1.50 [1.36, 1.64]
Boredom ~ condition	1.13 [0.89, 1.38] (BF > 500) *d* = 0.72 [0.54, 0.89]	0.69 [0.38, 1.00] (BF > 500) *d* = 0.43 [0.23, 0.63]	0.60 [0.31, 0.89] (BF = 198.00) *d* = 0.35 [0.17, 0.53]	0.74 [0.40, 1.08] (BF > 500) *d* = 0.41 [0.21, 0.61]	0.91 [0.75, 1.08] (BF > 500) *d* = 0.55 [0.45, 0.66]
Fatigue ~ condition	2.05 [1.84, 2.25] (BF > 500) *d* = 1.65 [1.37, 1.90]	1.41 [1.13, 1.69] (BF > 500) *d* = 0.97 [0.72, 1.21]	1.92 [1.65, 2.19] (BF > 500) *d* = 1.25 [1.01, 1.49]	1.98 [1.64, 2.31] (BF > 500) *d* = 1.27 [0.97, 1.60]	2.49 [2.35, 2.63] (BF > 500) *d* = 1.86 [1.70, 2.02]
Boundary ~ condition + congruency					
condition	−0.004 [−0.006, −0.002] (BF = 25.45) *d* = −0.22 [−0.34, −0.10]	−0.002 [−0.005, 0.001] (BF = 0.08) *d* = −0.08 [−0.24, 0.07]	−0.005 [−0.008, −0.003] (BF > 500) *d* = −0.32 [−0.46, −0.18]	−0.006 [−0.01, 0.00] (BF = 0.60) *d* = −0.13 [−0.26, 0.004]	−0.004 [−0.005, −0.002] (BF = 29.10) *d* = −0.15 [−0.22, −0.07]
congruency	−0.02 [−0.02, −0.02] (BF > 500) *d* = −1.12 [−1.25, −0.98]	−0.02 [−0.02, −0.01] (BF > 500) *d* = −0.94 [−1.12, −0.76]	−0.01 [−0.02, −0.01] (BF > 500) *d* = −0.88 [−1.04, −0.72]	−0.01 [−0.02, −0.004] (BF = 0.30) *d* = −0.27 [−0.44, −0.09]	−0.02 [−0.02, −0.01] (BF > 500) *d* = −0.68 [−0.75, −0.60]
Drift rate ~ condition + congruency					
condition	−0.007 [−0.01, −0.001] (BF = 0.44) *d* = −0.14 [−0.26, −0.02]	−0.01 [−0.02, −0.002] (BF = 0.89) *d* = −0.20 [−0.36, −0.05]	0.001 [−0.006, 0.009] (BF = 0.04) *d* = 0.03 [−0.11, 0.16]	−0.003 [−0.01, 0.007] (BF = 0.07) *d* = −0.05 [−0.21, 0.11]	−0.003 [−0.007, 0.001] (BF = 0.06) *d* = −0.05 [−0.13, 0.02]
congruency	−0.02 [−0.02, −0.02] (BF > 500) *d* = −1.12 [−1.25, −0.98]	−0.02 [−0.02, −0.01] (BF > 500) *d* = −0.94 [−1.12, −0.76]	−0.01 [−0.02, −0.01] (BF > 500) *d* = −0.88 [−1.04, −0.72]	−0.01 [−0.02, −0.004] (BF > 500) *d* = −0.27 [−0.44, −0.09]	−0.02 [−0.02, −0.01] (BF > 500) *d* = −0.68 [−0.75, −0.60]
Boundary ~ fatigue	−0.001 [−0.002, 0.00] (BF = 1.19) *d* = −0.09 [−0.15, −0.03]	0.00 [−0.001, 0.001] (BF = 0.02) *d* = 0.01 [−0.07, 0.09]	−0.002 [−0.003, −0.001] (BF = 317.55) *d* = −0.14 [−0.20, −0.08]	−0.002 [−0.004, 0.00] (BF = 0.10) *d* = −0.04 [−0.10, 0.02]	−0.001 [−0.002, 0.00] (BF = 1.88) *d* = −0.05 [−0.08, −0.02]
Boundary ~ frustration	−0.001 [−0.002, 0.00] (BF = 0.96) *d* = −0.08 [−0.14, −0.03]	0.00 [−0.001, 0.002] (BF = 0.02) *d* = 0.02 [−0.07, 0.11]	−0.002 [−0.003, −0.001] (BF = 6.86) *d* = −0.12 [−0.19, −0.05]	−0.002 [−0.005, 0.001] (BF = 0.10) *d* = −0.04 [−0.12, 0.03]	−0.001 [−0.002, 0.00] (BF = 0.25) *d* = −0.04 [−0.08, −0.009]
Boundary ~ boredom	−0.001 [−0.002, 0.00] (BF = 0.08) *d* = −0.06 [−0.13, 0.009]	0.00 [−0.002, 0.001] (BF = 0.03) *d* = 0.00 [−0.10, 0.09]	−0.001 [−0.002, 0.00] (BF = 0.17) *d* = −0.09 [−0.17, −0.005]	0.00 [−0.003, 0.004] (BF = 0.06) *d* = 0.005 [−0.08, 0.09]	0.00 [−0.001, 0.00] (BF = 0.02) *d* = −0.008 [−0.05, 0.04]
Drift rate ~ fatigue	−0.001 [−0.003, 0.001] (BF = 0.03) *d* = −0.03 [−0.09, 0.03]	−0.003 [−0.006, 0.00] (BF = 0.09) *d* = −0.07 [−0.15, 0.01]	0.00 [−0.002, 0.003] (BF = 0.02) *d* = 0.01 [−0.05, 0.07]	0.00 [−0.004, 0.003] (BF = 0.03) *d* = −0.007 [−0.07, 0.06]	0.00 [−0.002, 0.001] (BF = 0.01) *d* = −0.007 [−0.04, 0.02]
Drift rate ~ frustration	−0.002 [−0.004, 0.00] (BF = 0.23) *d* = −0.07 [−0.12, −0.01]	−0.002 [−0.006, 0.001] (BF = 0.06) *d* = −0.06 [−0.15, 0.03]	0.00 [−0.003, 0.003] (BF = 0.03) *d* = −0.006 [−0.07, 0.06]	−0.002 [−0.006, 0.003] (BF = 0.05) *d* = −0.03 [−0.11, 0.05]	−0.001 [−0.003, 0.00] (BF = 0.05) *d* = −0.03 [−0.07, 0.005]
Drift rate ~ boredom	−0.004 [−0.006, −0.001] (BF = 0.99) *d* = −0.10 [−0.17, −0.03]	−0.003 [−0.007, −0.001] (BF = 0.11) *d* = −0.08 [−0.18, 0.01]	−0.001 [−0.005, 0.002] (BF = 0.04) *d* = −0.03 [−0.11, 0.05]	−0.005 [−0.01, 0.00] (BF = 0.27) *d* = −0.09 [−0.18, 0.004]	−0.003 [−0.004, −0.001] (BF = 0.67) *d* = −0.06 [−0.11, −0.02]

Note. Numbers within brackets are the upper and lower limits of 95% highest posterior density intervals. Informed priors reflecting Cohen's *d* = 0.28 (*SD* = 0.14) were created by rescaling the expected effect size to the raw scale of each outcome measure. Bayes factors were computed using bridge sampling. Bayes factor > 1 indicates evidence for the experimental hypothesis, whereas values < 1 indicates evidence for the null hypothesis. BF = Bayes factor.

Table 2
Exploratory Analyses: Bayesian Multilevel Model Parameter Estimates Using Zero-Centered Normal Priors

	Study 1 (20 min)	Study 2 (15 min)	Study 3 (5 min)	Study 4 (10 min)	Overall
Boundary ~ demand	−0.001 [−0.002, 0.00] (BF = 0.65) *d* = −0.05 [−0.11, 0.002]	0.00 [−0.001, 0.00] (BF = 0.19) *d* = −0.03 [−0.10, 0.04]	−0.001 [−0.002, 0.00] (BF = 226.86) *d* = −0.10 [−0.16, −0.05]	−0.001 [−0.003, 0.001] (BF = 0.54) *d* = −0.03 [−0.09, 0.02]	−0.001 [−0.002, −0.001] (BF = 17.21) *d* = −0.05 [−0.08, −0.02]
Boundary ~ effort	0.00 [−0.001, 0.00] (BF = 0.30) *d* = −0.04 [−0.10, 0.01]	0.00 [−0.002, 0.00] (BF = 0.22) *d* = −0.03 [−0.10, 0.04]	−0.001 [−0.002, 0.00] (BF = 66.44) *d* = −0.10 [−0.15, −0.04]	−0.001 [−0.004, 0.00] (BF = 0.57) *d* = −0.03 [−0.09, 0.02]	−0.001 [−0.002, 0.00] (BF = 16.04) *d* = −0.05 [−0.08, −0.02]
Drift rate ~ demand	0.00 [−0.002, 0.002] (BF = 0.13) *d* = −0.01 [−0.07, 0.05]	−0.002 [−0.005, 0.00] (BF = 0.72) *d* = −0.06 [−0.13, 0.006]	0.00 [−0.002, 0.003] (BF = 0.17) *d* = 0.02 [−0.04, 0.07]	0.00 [−0.003, 0.003] (BF = 0.20) *d* = 0.008 [−0.05, 0.06]	0.00 [−0.001, 0.001] (BF = 0.08) *d* = −0.005 [−0.03, 0.02]
Drift rate ~ effort	−0.001 [−0.003, 0.001] (BF = 0.23) *d* = −0.03 [−0.08, 0.03]	−0.002 [−0.005, 0.00] (BF = 0.62) *d* = −0.06 [−0.14, 0.01]	0.00 [−0.002, 0.002] (BF = 0.13) *d* = 0.00 [−0.05, 0.05]	0.00 [−0.003, 0.004] (BF = 0.20) *d* = 0.01 [−0.05, 0.07]	0.00 [−0.002, 0.00] (BF = 0.10) *d* = −0.01 [−0.04, 0.02]
Boundary ~ condition * congruency	0.003 [−0.001, 0.006] (BF = 1.04) *d* = 0.14 [−0.08, 0.36]	−0.001 [−0.006, 0.004] (BF = 0.64) *d* = −0.06 [−0.34, 0.21]	0.001 [−0.003, 0.005] (BF = 0.58) *d* = 0.08 [−0.18, 0.33]	0.00 [−0.008, 0.007] (BF = 0.89) *d* = −0.01 [−0.18, 0.16]	0.00 [−0.003, 0.004] (BF = 0.43) *d* = 0.03 [−0.11, 0.17]
Drift rate ~ condition * congruency	0.002 [−0.009, 0.01] (BF = 0.47) *d* = 0.04 [−0.18, 0.26]	0.001 [−0.01, 0.01] (BF = 0.58) *d* = 0.02 [−0.26, 0.30]	−0.003 [−0.02, 0.01] (BF = 0.58) *d* = −0.05 [−0.30, 0.19]	0.003 [−0.01, 0.02] (BF = 0.74) *d* = 0.04 [−0.21, 0.30]	0.001 [−0.007, 0.009] (BF = 0.32) *d* = 0.02 [−0.12, 0.16]
Stroop ACC ~ condition * congruency					
condition	−0.007 [−0.02, 0.00] (BF = 1.73) *d* = −0.11 [−0.22, 0.009]	−0.003 [−0.01, 0.007] (BF = 0.49) *d* = −0.04 [−0.19, 0.11]	−0.003 [−0.01, 0.006] (BF = 0.46) *d* = −0.05 [−0.18, 0.09]	−0.004 [−0.01, 0.007] (BF = 0.58) *d* = −0.05 [−0.21, 0.10]	−0.005 [−0.01, 0.00] (BF = 1.21) *d* = −0.08 [−0.15, −0.01]
congruency	−0.09 [−0.10, −0.08] (BF > 500) *d* = −1.28 [−1.41, −1.14]	−0.09 [−0.10, −0.08] (BF > 500) *d* = −1.36 [−1.56, −1.17]	−0.07 [−0.08, −0.06] (BF > 500) *d* = −0.99 [−1.15, −0.83]	−0.07 [−0.08, −0.05] (BF > 500) *d* = −0.95 [−1.15, −0.76]	−0.08 [−0.08, −0.07] (BF > 500) *d* = −1.16 [−1.24, −1.08]
condition * congruency	−0.001 [−0.02, 0.01] (BF = 0.56) *d* = −0.02 [−0.22, 0.18]	0.00 [−0.02, 0.02] (BF = 0.68) *d* = −0.005 [−0.25, 0.25]	−0.003 [−0.02, 0.01] (BF = 0.68) *d* = −0.05 [−0.28, 0.18]	−0.001 [−0.02, 0.02] (BF = 0.71s) *d* = −0.02 [−0.27, 0.22]	−0.002 [−0.01, 0.007] (BF = 0.35) *d* = −0.03 [−0.17, 0.10]
Stroop RT ~ condition * congruency					
condition	−0.004 [−0.01, 0.003] (BF = 0.42) *d* = −0.07 [−0.19, 0.05]	0.009 [−0.002, 0.02] (BF = 1.07) *d* = 0.14 [−0.02, 0.30]	−0.01 [−0.02, −0.003] (BF = 7.15) *d* = −0.19 [−0.33, −0.05]	−0.01 [−0.02, −0.004] (BF = 15.03) *d* = −0.25 [−0.42, −0.08]	−0.006 [−0.01, −0.001] (BF = 2.32) *d* = −0.09 [−0.17, −0.02]
congruency	0.10 [0.10, 0.11] (BF > 500) *d* = 1.72 [1.57, 1.87]	0.11 [0.10, 0.12] (BF > 500) *d* = 1.67 [1.46, 1.88]	0.10 [0.09, 0.11] (BF > 500) *d* = 1.61 [1.43, 1.78]	0.12 [0.11, 0.13] (BF > 500) *d* = 2.05 [1.80, 2.27]	0.11 [0.10, 0.11] (BF > 500) *d* = 1.73 [1.63, 1.82]
condition * congruency	−0.006 [−0.02, 0.008] (BF = 0.56) *d* = −0.10 [−0.32, 0.13]	−0.001 [−0.02, 0.02] (BF = 0.52) *d* = −0.02 [−0.30, 0.28]	−0.004 [−0.02, 0.01] (BF = 0.52) *d* = −0.07 [−0.32, 0.19]	−0.007 [−0.02, 0.01] (BF = 0.66) *d* = −0.12 [−0.43, 0.19]	−0.005 [−0.01, 0.003] (BF = 0.55) *d* = −0.09 [−0.23, 0.05]

Note. Numbers within brackets are the upper and lower limits of 95% highest posterior density intervals. Bayes factor > 1 indicates evidence for the experimental hypothesis, whereas values < 1 indicates evidence for the null hypothesis. Bayes factors were computed using bridge sampling. BF = Bayes factor. RT = Stroop task reaction time. ACC: Stroop task accuracy.

Discussion

Our results provide insights into the psychology of effort exertion. Across four studies, our high (vs. low) demand manipulation was highly "depleting" as it robustly elicited strong effort and fatigue sensations. Diffusion model analyses provided insights into the effects of effort exertion on cognitive processes that have been previously unexamined.

Of the two preregistered effects on the latent parameters, Bayesian analyses provided strong evidence for reduced boundary but not drift rate after participants completed the high (vs. low) demand task. The lack of evidence for reduced drift rate suggests effort exertion did not worsen participants' subsequent task performance or their abilities to process information. However, reduced boundary separation suggests participants responded less cautiously, as if they cared less about the task and had lost some of their "will" to persist and engage fully.

Crucially, the reduced-boundary effect was not limited to situations involving inhibition (i.e., incongruent Stroop trials), consistent with our theoretical position that exerting effort leads to task re-prioritization and disengagement with ongoing tasks (Inzlicht et al., 2014). Accordingly, depletion should impair performance on incongruent *and* congruent Stroop trials. Indeed, exploratory analyses revealed that overall Stroop reaction time and accuracy were reduced in the high (vs. low) demand condition, though the effect was much weaker relative to the boundary effect. Furthermore, results from an extended diffusion model for conflict tasks also suggested effort exertion affected only boundary separation but not controlled or automatic information integration processes.

Further evidence for our theoretical view comes from the finding that participants reported increased boredom in the high (vs. low) demand condition, which might indicate unsuccessful attentional engagement when people feel either unable or unwilling to engage with ongoing tasks (Westgate & Wilson, 2018). Moreover, participants who reported increased fatigue, demand, or effort also had reduced boundary parameters, but these exploratory effects should be interpreted with caution.

Our findings suggest that even when tasks elicit strong subjective states related to fatigue, traditional behavioral measures might lack sensitivity to detect downstream effects. Instead, latent variables might be more sensitive. For example, the reduced-boundary effect was about twice as large as the reduced overall Stroop reaction time and accuracy effects, likely because the diffusion model solves the speed-accuracy trade-off associated with reaction-time tasks. Given the strengths of the diffusion model (see Evans & Wagenmakers, 2019), we suggest other researchers apply similar approaches or reanalyze previous depletion studies that used speeded reaction-time tasks.

Depletion proponents might celebrate because our results provide strong evidence for and further insights into depletion effects, as well as against the null hypothesis that depletion effects do not exist. Skeptics, however, will hasten to highlight various limitations. Only one of two hypotheses were confirmed—and only meta-analytically, with merely two of four individual studies providing evidence for our preregistered hypotheses. Nevertheless, the small but meaningful effect size (reduced boundary effect $d = -0.13$) suggests researchers hoping to examine similar effects should use within-subjects to ensure sufficient statistical power, especially since the boundary effect was present even after accounting for within-subjects learning effects in our studies. Despite these issues, our work has numerous strengths—strong manipulations, preregistered hypotheses, and cognitive modelling—that have allowed us to rigorously examine the cognitive processes underlying effort exertion.

Conclusion

Our paradigm robustly elicited feelings such as effort and fatigue, highlighting its utility for studying these subjective states. Bayesian analyses provided strong evidence for the idea that people disengage after exerting effort. Although we failed to find support for all our hypotheses, we have learned that laboratory depletion effects are elusive even with strong manipulations and latent variables that capture meaningful cognitive processes. But our rigorous approach has much potential to facilitate future empirical and theoretical developments.

Author Contributions

H. Lin, B. Saunders, M. Friese, and M. Inzlicht developed the study concept and contributed to the study design. H. Lin performed testing and data collection. H. Lin and N. J. Evans performed the data analysis and interpretation under the supervision of B. Saunders, M. Friese, N. J. Evans, and M. Inzlicht. H. Lin drafted the manuscript, and the remaining authors provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgements

We thank Colin Kupitz, Joachim Vandekerckhove, and Shravan Vasishth for their guidance, and Eric-Jan Wagenmakers and another reviewer for their feedback.

References

Baumeister, R. F., & Vohs, K. D. (2016). Strength model of self-regulation as limited resource: Assessment, controversies, update. In *Advances in Experimental Social Psychology* (pp. 67-127). Elsevier. doi:10.1016/bs.aesp.2016.04.001

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. doi:10.1037/0022-3514.74.5.1252

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28. doi:10.18637/jss.v080.i01

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796-815. doi:10.1037/xge0000083

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*, 115-144. doi:10.1177/2515245919847196

Dai, H., Milkman, K. L., Hofmann, D. A., & Staats, B. R. (2015). The impact of time at work and time off from work on rule compliance: The case of hand hygiene in health care. *Journal of Applied Psychology*, *100*(3), 846-862. doi:10.1037/a0038067

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. doi:10.3389/fpsyg.2014.00781

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., . . . Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051-1069. doi:10.3758/s13423-017-1417-2

Dutilh, G., Krypotos, A. M., & Wagenmakers, E. J. (2011). Task-related versus stimulus-specific practice. *Experimental Psychology*, *58*(6), 434-442. doi:10.1027/1618-3169/a000111

Etherton, J. L., Osborne, R., Stephenson, K., Grace, M., Jones, C., & De Nadai, A. S. (2018). Bayesian analysis of multimethod ego-depletion studies favours the null hypothesis. *British Journal of Social Psychology*, *57*(2), 367-385. doi:10.1111/bjso.12236

Evans, N. J., & Wagenmakers, E.-J. (2019). Evidence accumulation models: Current limitations and future directions. Retrieved from https://psyarxiv.com/74df9/download?format=pdf

Evans, N. J., & Servant, M. (2019). A comparison of conflict diffusion models in the flanker task through pseudolikelihood Bayes factors. *Psychological Review*. doi:10.1037/rev0000165

Francis, Z., & Job, V. (2018). Lay theories of willpower. *Social and Personality Psychology Compass*, *12*(4), e12381. doi:10.1111/spc3.12381

Francis, Z., Milyavskaya, M., Lin, H., & Inzlicht, M. (2018). Development of a within-subject, repeated-measures ego-depletion paradigm. *Social Psychology*, *49*(5), 271-286. doi:10.1027/1864-9335/a000348

Friese, M., & Frankenbach, J. (2019). p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*. doi:10.1037/met0000246

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review*, *23*(2), 107-131. doi:10.1177/1088868318762183

Garavan, H., Ross, T. J., Li, S.-J., & Stein, E. A. (2000). A parametric manipulation of central executive functioning. *Cerebral Cortex*, *10*(6), 585-592. doi:10.1093/cercor/10.6.585

Garrison, K. E., Finley, A. J., & Schmeichel, B. J. (2019). Ego depletion reduces attention control: Evidence from two high-powered preregistered experiments. *Personality and Social Psychology Bulletin*, *45*(5), 728-739. doi:10.1177/0146167218796473

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546-573. doi: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*(4), 495-525. doi:10.1037/a0019486

Hirshleifer, D., Levi, Y., Lourie, B., & Teoh, S. H. (2019). Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics*, *133*(1), 83-98. doi:10.1016/j.jfineco.2019.01.005

Hockey, R. (2013). *The psychology of fatigue: Work, effort and control*. Cambridge University Press.

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, *18*(3), 127-133. doi:10.1016/j.tics.2013.12.009

Inzlicht, M., & Friese, M. (2019). The past, present, and future of ego depletion. *Social Psychology*, *50*(5-6), 370-378. doi:10.1027/1864-9335/a000398

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course*. New York, NY: Cambridge University Press.

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1). doi:10.5334/irsp.289

Lin, H. (2019). hauselin/hausekeep: third release (Version v0.0.0.9003-alpha). *Zenodo*. doi:10.5281/zenodo.2555874

Milyavskaya, M., Inzlicht, M., Johnson, T., & Larson, M. J. (2019). Reward sensitivity following boredom and cognitive effort: A high-powered neurophysiological investigation. *Neuropsychologia*, *123*, 159-168. doi:10.1016/j.neuropsychologia.2018.03.033

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8-14. doi:10.1177/0963721411429458

Moller, A. C., Deci, E. L., & Ryan, R. M. (2006). Choice and ego-depletion: The moderating role of autonomy. *Personality and Social Psychology Bulletin*, *32*(8), 1024-1036. doi:10.1177/0146167206288008

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873-922. doi:10.1162/neco.2008.12-06-420

Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *Journal of Experimental Psychology: General*, *136*(2), 241-255. doi:10.1037/0096-3445.136.2.241

Sjåstad, H., & Baumeister, R. F. (2018). The future and the will: Planning requires self-control, and ego depletion leads to planning aversion. *Journal of Experimental Social Psychology*, *76*, 127-141. doi:10.1016/j.jesp.2018.01.005

Tuk, M. A., Zhang, K., & Sweldens, S. (2015). The propagation of self-control: Self-control in one domain simultaneously improves self-control in other domains. *Journal of Experimental Psychology: General*, *144*(3), 639-654. doi:10.1037/xge0000065

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148-174. doi:10.1016/j.cogpsych.2015.02.005

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*(2), 547-556. doi:10.3758/s13423-016-1081-y

Wagenmakers, E. J., & Gronau, Q. (2017). Redefine statistical significance XIII: The case of ego depletion. Retrieved Dec 28, 2017 from https://www.bayesianspectacles.org/redefine-statistical-significance-xiii-the-case-of-ego-depletion/

Wagenmakers, E. J., van der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3-22. doi:10.3758/BF03194023

Westfall, J. (2015). PANGEA: Power ANalysis for GEneral Anova designs. *Unpublished manuscript Available at http://jakewestfall org/publications/pangea pdf*. Retrieved from http://jakewestfall.org/publications/pangea.pdf

Westgate, E. C., & Wilson, T. D. (2018). Boring thoughts and bored minds: The MAC model of boredom and cognitive engagement. *Psychological Review*, *125*(5), 689-713. doi:10.1037/rev0000097