Supplemental Materials for

Development of a Within-Subject, Repeated-Measures Ego Depletion Paradigm: Inconsistent Results and Future Recommendations

Table of Contents

Additional Methods

Undergraduate participants in Studies 1, 3, and 5 were from the University of Toronto Scarborough, and participants in Study 6 were from the University of Sussex. For in-lab studies at University of Toronto, up to five participants were seated at separated desks with individual computers, and conducted the studies programmed in MediaLab and DirectRT. Online studies were conducted with Qualtrics (experiment files available at https://osf.io/rh2gv/).

All selected behavioural self-control measures were modified to fit into a repeated-measures design by ensuring that the measure took no longer than two minutes (and generally much less than that). We attempted to select measures that would not be susceptible to practice effects or to mood inductions. Three of the dependent variables have been previously used as dependent measures in depletion paradigms (CET in Schmeichel et al., 2003; solvable anagrams in Baumeister et al., 1998; inhibition reaction time tasks in Inzlicht & Gutsell, 2007). The remaining two studies use the anchoring effect heuristic, which, to our knowledge, had not been previously used in depletion experiments.

Depleting Manipulation: The Add-3 Task

Participants are shown four random numbers for one second each, and must store the numbers in working memory and add three to each of the numbers (1 becomes 4, 9 becomes 2, etc.), before typing in their answer. They then have a four-second response window to type their answer. This task is extremely effortful, as measured both by subjective report of effort (and, seen below, self-reports of fatigue) and large pupil dilation while performing the task (Kahneman et al., 1969), making it a good candidate to induce depletion. Although there is no explicit inhibitory component to this task, cognitive control is required to maintain items in working memory (e.g., Schmeichel, 2007) and to stay focused and persevere on the task. Similarly effortful tasks, such as sustained attention tasks, also cause decreasing performance such that the more effortful the task, the greater the performance decrement – these decrements are explained using a similar resource model to the original model of depletion (Pattyn, Neyt, Henderickx, & Soetens, 2008; Thomson, Besner, & Smilek, 2015).

Self-Reported Mood and Energy

To indicate mood and energy levels, participants were presented with numberless scales modelled after visual analogue scales (Crichton, 2001); the mood scale was anchored with ‘pleasant’ at the top (or right) and ‘unpleasant’ at the bottom (or left) and the energy scale anchored with ‘energized’ and ‘fatigued’, and participants chose the point on the line which most closely represented their current state. These mood scales went from 1-100 for online participants, and 1-12 for in-lab participants (transformed to 1-100 for consistent analyses).

Cognitive Estimation Task (CET)

Depleted participants have previously performed more poorly on the CET, theoretically because participants who are unable or unwilling to self-regulate do not appropriately monitor the quality of their answers and are more likely to write down the first guess that comes to mind, regardless of the quality. Performance on the CET does not depend on specific crystallized knowledge, due to the ambiguity of the questions (Della Sala, MacPherson, Phillips, Sacco, & Spinnler, 2004).

Five studies used the same set of CET questions, which were first piloted online on Mechanical Turk (*N* = 138) to ensure each question had a unimodal distribution, and to collect reference means and standard deviations to each questions. These parameters were later used to transform experimental participant’s raw answers into z-scores, to facilitate comparisons between questions. To determine a participant’s self-control score for a given block, we first translated each of the three CET answers into absolute z-scores, using the mean and standard deviations from reference pilot data. For example, the mean piloted response to, “In what year was the O’Henry chocolate bar released?” was 1923 (*SD* =28). A participant who answered ‘1895’ would receive an absolute z-score of 1 for that question. Based on the pilot data distributions, 25 of the 67 questions (e.g., “How many babies are born in Canada each year?”) were also log-transformed before conversion to z-scores due to answers commonly stretching across multiple orders of magnitude (e.g., 10 000 to 1 000 000). To reduce the effect of outliers, participant answers with high absolute z-scores (those above the 95th quantile, for each study) were converted to the value of the 95th quantile. Then the absolute z-scores for the three estimation questions (or four questions, for Study 10) within a single block were averaged. The same process was used on a different set of questions for Study 6.

Flanker Task

There are two types of trials: compatible trials present the target letter surrounded by four flanking letters of the same type (e.g., SSSSS) and incongruent trials present the target letter surrounded by the opposite letter (e.g., SSHSS). Incongruent trials are more difficult, are associated with longer reaction times and more errors, and require greater inhibitory control of pre-potent responses (Eriksen & Eriksen, 1974; van Steenbergen, Band, & Hommel, 2009). For each flanker trial, the participant focuses on a fixation cross before the trial begins, then the flanking letters appear on the screen alone for 100ms, and lastly the central target letter appears with the flanker for 250ms. Participants must suppress their response to the flanking letters to appropriately respond to the target by pressing the associated button within the 1500s response window.

Performance on the flanker task can be analyzed a number of ways, including by examining error rates or reaction times. As our primary dependent measure (used in the meta-analyses), we measured the number of errors on incongruent trials. More errors on incongruent trials is a sign of poorer performance, particularly when not accompanied by increased reaction times that can signal a time-accuracy tradeoff. Because of the limited response window and a relatively high proportion of errors (15.5% of incongruent trials), error rate was a viable measure of performance. As a secondary dependent measures, we analyzed the reaction times for incongruent trials, controlling for reaction time on congruent trials. Although not included in the aggregated results tables or meta-analyses, results with reaction time are available here, below (in Supplemental materials).

Anagrams Task

After the depleting or rejuvenating manipulation, participants were given 90 seconds to solve as many anagrams as they could. It should be noted that – unlike in previous studies on solvable anagrams (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Gordijn, Hindriks, Koomen, Dijksterhuis, & Van Knippenberg, 2004) – participants could only see a single anagram at a time. This differs from other published solvable anagram tasks, where participants are given 12 to 25 anagrams on a piece of paper and have between five and twenty minutes to solve as many as they can.

For the aggregated result tables and the meta-analyses, we presented the number of correctly solved anagrams as the primary dependent variable. This measure of performance is commonly used (e.g., Baumeister et al., 1998; Boucher & Kofos, 2012) and is the most face-valid measure of performance for the solvable anagrams task. Other variables – such as time spent per anagram – may diverge based on different participants' strategies towards the tasks (e.g., seeing many anagrams and solving only the easy ones quickly, versus seeing few anagrams and solving each of them slowly), and thus may not linearly map on to performance. Results with these secondary dependent variables are available below, in the Supplemental Materials.

A single five-letter scrambled anagram was presented on the screen, and the participant could choose to answer the anagram or to press a ‘SKIP’ button in the bottom corner of the screen to move to the next anagram. Once a particular anagram was skipped, participants could not return to see the same anagram again. Participants would continue to work on the anagrams, one at a time, until the 90 seconds ran out and the experiment automatically progressed. For each separate block, we then computed separate variables for the number of correctly solved anagrams, incorrectly solved anagrams, skipped anagrams, as well as the total number seen by the participant that block. We also calculated the average amount of time taken to solve an anagram correctly, and the average amount of time taken before skipping an anagram.

Anchoring Task

To measure the anchoring effect, we modified the questions used in the CET to include an anchor value. Participants were first asked whether the true answer was higher or lower than the anchor value, and then responded with their own open-ended answer. Provided anchors were defined as the value at the 10th or 90th percentile of the cleaned pilot reference dataset (with outliers removed), which was generally equivalent to the 15th or 85th percentile of the full dataset (Jacowitz & Kahneman, 1995; Strack & Mussweiler, 1997). Anchors were rounded to the nearest round number (e.g., 42.3% rounded to 40%). The questions were evenly split between high and low anchors, and the three questions of a single block included both high and low anchors. Participant’s estimates were transformed into the degree of anchor effect (%) according to the following (Jacowitz & Kahneman, 1995):

$$\frac{\text{Participant's estimate} - \text{Mean estimate}}{\text{Provided anchor} - \text{Mean estimate}}$$

Thus, someone who guessed that the anchor was the true value would receive a score of 1, someone who responded half-way between the provided anchor and the average baseline response would receive a 0.5, and someone who responded with precisely the average baseline response (not influenced by the anchor) would receive a 0.

Pre-registrations

The online pre-registration for Study 6, including hypotheses and detailed methods, is available at osf.io/2iwk5. The pre-registration for Study 8a and 8b – including experiment files, sample size, and hypotheses – is available at https://osf.io/7tm9j/. Other studies were not pre-registered. More broadly, materials and analysis code are available at the primary OSF page for this project here: https://osf.io/rh2gv/

Additional Results

Condition by Trial Interaction Predicting Self-Reported State

In no individual study did condition and block significantly interact to predict mood or energy. Condition and block thus appear to be generally distinct predictors of self-reported subjective state. However, when analyzing data from all of the depletion vs. recovery studies together (using three-level hierarchical models; blocks nested within participant nested within study), we find that condition and block do interact to predict energy (*F*(1, 5880) = 7.90, *p* = .005); the effect of the depletion manipulation on energy is generally smaller at the beginning of the paradigm (*B* = 4.75, *SE* = .60, *t*(5880) = 7.93, *p* < .001), compared to after six blocks (*B* = 5.82, *SE* = 0.38, t(5880) = 15.25, *p* < .001). The same was not true for mood. Mood continued to be individually predicted by the condition manipulation and the block, without those factors interacting (*F*(1, 5885) = 0.20, *p* = .66).

Moderation by Individual Difference Measures

This model included the same two random effects as other models (participant and question set) plus five fixed effects: condition, block, the moderator (e.g. willpower belief), the moderator-by-condition interaction, and the moderator-by-block interaction.

Implicit Willpower Theories. Participants in Study 1 completed the ‘Implicit Beliefs of Willpower’ scale (Job, Dweck, & Walton, 2010), a six-item questionnaire that measures whether participants believe that doing strenuous activities causes their willpower to be depleted (a limited resource view) or whether difficult activities energize them (a non-limited view). This questionnaire has been previously found to moderate the effectiveness of depletion manipulations, with people who hold a limited resource view showing a greater decrease in performance on the second self-control task (Job et al., 2010). We were interested in whether these moderation results would replicate in this repeated-measures paradigm. Originally, we predicted that willpower beliefs would moderate the efficacy of the depletion manipulation on both self-reported fatigue and on CET performance, such that limited theorists would show greater differences between the recovery and depletion conditions compared to non-limited theorists. After observing significant block effects in the paradigm, where participants generally report more fatigue across the course of the experiment, we would also predict that the accumulation of fatigue and worsening mood would be exacerbated for limited theorists, compared to non-limited theorists.

Lay beliefs in willpower did not impact performance on the CET task directly (*F*(1, 827) = 1.78, *p* = .18) , nor did it moderate the effect of condition manipulation (*F*(1, 827) = 0.70, p =

.40) or of block (*F*(1, 827) = 0.02, *p* = .89) on cognitive estimation scores. However, willpower theories did significantly moderate some of the changes in people's self-reported mood and energy. In particular, while participants generally reported becoming more fatigued over the course of the study, this decrease in energy was more pronounced in limited theorists (at +1 *SD*, *B* = 1.07, *SE* = 0.13 per block) compared to non-limited theorists (at -1 *SD*, *B* = 0.51, *SE* = 0.13; interaction *F*(1, 916) = 9.04, *p* = .003). Additionally, while participants generally reported more negative mood after the depletion manipulation compared to after the recovery manipulation, this different was also more pronounced in limited theorists (*B* = 6.84, *SE* = 0.85) compared to non-limited theorists (*B* = 3.68, *SE* = 0.85; interaction *F*(1, 916) = 8.16 , *p* = .004). Inconsistently, however, neither the difference in self-reported energy between conditions nor the gradual decrease in mood across the course of the study was significantly moderated by willpower theory (*p* > .22).

As expected, there was no main effect of willpower theories on energy levels (*F*(1, 922) = 1.77, *p* = .18) or mood (*F*(1, 922) = .03, *p* = 1.77).

Enjoyment of Videos. In Study 5 (in-lab, DV: solvable anagrams), participants indicated how enjoyable they generally found the short videos. People who reported liking the videos more experienced less accumulation of fatigue and less decrease in mood across the course of the experiment than the people who liked the videos less (fatigue *F*(1, 952) = 9.53, *p* = .002, mood *F*(1, 952) = 11.67, *p* <.001). Those who particularly enjoyed the videos (one *SD* above the mean) did not report significantly decreasing mood across the course of the experiment (*B* = -0.21, *SE* = 0.16, *t*(952) = 1.32, *p* = .19), while those who did not enjoy the videos (one *SD* below the mean) experienced a decrease in mood three times larger (*B* = - 0.72, *SE* = 0.16, *t*(952) =

4.53, p < . 0001). Enjoyment of the videos did not moderate the manipulation condition effect's on either mood or fatigue.

Secondary Dependent Variables

The flanker and the anagram tasks had other potential dependent variables. While we selected the number of errors (for incongruent trials) as the primary dependent variable for the flanker task (Study 3), and the number of anagrams solved as the primary dependent variable for the anagram task (Study 5), we can also examine reaction time for both tasks. For the flanker task, there was no condition effect on the reaction time for correctly answered incongruent trials (controlling for reaction time of correct congruent trials; B = 1.49, t(622) = 0.98, p = .32) although participants did have shorter reaction times to the incongruent trials later in the experiment (B = -1.08, t(622) = 3.58, p < .001). Reaction time was not predicted by the interaction between block and condition, however (B = .49, t(622) = 1.65, p = .099; simple effect of condition at block 1, t(622) = 0.83, p = .41). This suggests that the general increase in error rates across the course of the flanker experiment (the block effect) may have been driven by a change in participants' speed-accuracy trade-off (Heitz, 2014). However, the interaction between the condition and block effects on the number of incongruent errors – the higher rate of errors after the depleting manipulation early in the experiment – cannot be explained by changes in the speed-accuracy trade-off.

In Study 5, we found a significant main effect of condition on perseverance (time before skipping an unsolved anagram), although not as predicted. We had originally predicted more self-control to be associated with more persistence on each anagram (as in Baumeister et al., 1998). Instead, participants were significantly slower to skip anagrams (e.g. persevered longer) after depletion manipulations (B = 0.42 seconds, p = .02). With further exploratory analysis, however, we found that more so-called persistence was associated with fewer anagrams being solved. Because only one anagram was shown on the screen at a time, the optimal strategy for this task involved quickly determining whether an anagram was easily solvable and skipping to the next anagram if it seemed too difficult. There was thus a significant indirect effect of depletion manipulation on the number of anagrams solved (95% *CI* [-.34 to -.04]), although there was no significant direct effect of depletion on anagrams solved. Due to the exploratory nature of this analysis, this effect is not included in the aggregated results above. However, this result is consistent with findings by Aspinwall and Richter (1999) demonstrating that, when alternatives are available, higher mastery and optimism are associated with earlier disengagement with unsolvable (or seemingly unsolvable) anagrams.

References

Aspinwall, L. G., & Richter, L. (1999). Optimism and self-mastery predict more rapid disengagement from unsolvable tasks in the presence of alternatives. *Motivation and Emotion*, *23*(221–245), 1999.

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), 1252–65.

Gordijn, E. H., Hindriks, I., Koomen, W., Dijksterhuis, A., & Van Knippenberg, A. (2004). Consequences of stereotype suppression and internal suppression motivation: a self-regulation approach. *Personality & Social Psychology Bulletin*, *30*(2), 212–24. http://doi.org/10.1177/0146167203259935

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*(8 JUN), 1–19. http://doi.org/10.3389/fnins.2014.00150

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166. http://doi.org/10.1177/01461672952111004

Job, V., Dweck, C., & 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. http://doi.org/10.1177/0956797610384745

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*(3), 437–446.