META-ANALYTIC AND EMPIRICAL ESTIMATES OF THE RESOURCE DEPLETION EFFECT SIZE

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Tyler Andrew Yost

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chad J. Marsolek, Advisor

May, 2016

Abstract

Published empirical studies of self-control in humans have provided evidence suggesting that the ability to exert self-control relies on a limited resource. Recent failed replications of the resource depletion effect, in addition to conflicting meta-analytic evidence, have called the robustness of the resource depletion effect into question. This dissertation aims to obtain a more accurate estimate of the depletion effect size using both an empirical replication and a novel meta-analytic method, *p*-curve. Monte Carlo simulations testing the accuracy of *p*-curve effect size estimates in the presence of publication bias and questionable research practices are also reported. Simulation results show that *p*-curve effect size estimates are unaffected by publication bias but fluctuate wildly when questionable research practices are simulated. Results from the empirical replication and meta-analysis suggest that the resource depletion effect is not as robust as previously thought. Further work is necessary to reliably differentiate resource depletion effects from type I error.

Table of Contents

List of Tables iii

List of Figures iv

Chapter 1: Resource Depletion 1

Chapter 2: Behavioral Experiment 8

Chapter 3: Meta-Analysis and Publication Bias 20

Chapter 4: *p*-Curve Meta-Analysis 36

Chapter 5: *p*-Curve Monte Carlo Simulations 45

Chapter 6: Conclusion 51

References 57

Appendix A: Behavioral Experiment Wordlists 71

Appendix B: Behavioral Experiment Instructions 72

Appendix C: Literature Search Details 73

Appendix D: Abbreviated *p*-curve Disclosure Table 75

List of Tables

Table 1: Behavioral Experiment RT Data 15

Table 2: Behavioral Experiment Accuracy Data 15

Table 3: Simulated Questionable Research Practices 46

List of Figures

Figure 1: Behavioral Experiment Stimuli and Trial Timing....11

Figure 2: Behavioral Experiment RT Results....14

Figure 3: Behavioral Experiment Post-Hoc Power....17

Figure 4: Behavioral Experiment Bootstrapped Confidence Interval....19

Figure 5: Effect of Publication Bias on Meta-Analysis....25

Figure 6: Funnel Plot Example....26

Figure 7: Example *p*-curves....30

Figure 8: Example *pp*-curves....33

Figure 9: Literature Search Flowchart....39

Figure 10: *p*-Curve Results: HWSC Dataset....42

Figure 11: *p*-Curve Results: post-HWSC Dataset....43

Figure 12: *p*-Curve Results: Combined Datasets....44

Figure 13: *p*-Curve Simulation Results....49

Figure 14: Forest Plot....55

Chapter One: Resource Depletion

Why is it sometimes so difficult to exert mental effort or control our attention and emotions? The dynamics of controlling attention, or suppressing emotion, have been topics of inquiry since the work of James (1890) and Freud (1920). Recently, studies of self-control exertion have demonstrated that self-control appears to rely on an exhaustible resource; individuals who exert self-control use up these resources, and as a result, have greater difficulty exerting self-control on subsequent tasks. These resource depletion effects appear to have robust empirical support; however, recent empirical studies and advanced meta-analyses have suggested that depletion effects may be much weaker than previously thought.

Resource Depletion: Background

In a review of the literature on self-control and the circumstances under which self-control efforts fail, Baumeister and Heatherton (1996) concluded that self-control appears to be dependent on a limited resource that can be depleted after exertion. In order to test this theory, Muraven, Tice, and Baumeister (1998) conducted several experiments to demonstrate that effortful exertion of self-control impaired subsequent attempts at self-control. In one experiment, subjects squeezed a handgrip exerciser for as long as possible as a baseline measurement. Following this, subjects watched a brief video clip depicting environmental disasters. One group of subjects was told to suppress any emotions they felt while watching the video, while controls were given no particular instructions. Following the video, all subjects performed the handgrip task again as a dependent measure. Subjects who were instructed to suppress their emotions during the video did

not persist at the final handgrip task as long as controls, suggesting that suppressing emotion depleted subjects' self-regulatory resources.

Similar results were obtained by Baumeister, Bratslavsky, Muraven, and Tice (1998). In one experiment, subjects watched brief emotional video clips (some sad, some humorous), followed by performance of a difficult (but solvable) anagram task. One group of subjects was instructed to suppress their affective reactions to the video clip while watching it (e.g., maintain a neutral facial expression), while a control group of subjects were told to simply watch the video and to let their emotions flow. Subjects in the self-control group solved fewer anagrams after controlling their affective reactions to the video compared with controls who did not suppress affect, indicating that exertion of self-control impairs performance on subsequent difficult self-control tasks.

The experimental designs described above are frequently employed in resource depletion experiments and recur throughout this literature. Typically, subjects are divided into two groups, with each group performing easy or difficult versions of an initial task. Following the initial task, both groups perform a second task that requires self-control. Resource depletion occurs when subjects who initially exerted self-control perform more poorly on the dependent measure in the second task compared with subjects who did not initially exert self-control. Alternatively, some experiments use an A-B-A design, in which subjects first perform a difficult self-control task to obtain a baseline measure, then perform a difficult (or easy) self-control task, and then perform the initial task again to obtain a measure of impaired performance when the second task is difficult (e.g., Shamosh & Gray, 2007). Most of the experiments described in this first section utilize

one of these designs (or slight variations thereof) with various depleting tasks and dependent measures.

Resource depletion effects are not limited to situations in which suppression of or tolerance for affective responses is engaged in the depleting task. Purely cognitive tasks have been used to elicit resource depletion effects. Schmeichel (2007) had subjects perform a working memory task followed by suppressing affect elicited by a disturbing video clip. The experimental group performed a difficult working memory task, while controls were given a less-demanding version. Subjects' facial expressions were videotaped while watching the disturbing video, and they were later rated according to facial expression. Subjects who initially performed the difficult task exhibited more emotion compared to controls, reflecting impaired self-control. In a separate experiment, Schmeichel (2007) first asked subjects to view an emotionally disturbing video clip; some subjects were told to exaggerate their responses, while controls were instructed to simply watch the video. After the video, all subjects performed a difficult operation span task. Subjects who exaggerated expressions performed more poorly than controls on the operation span task, demonstrating that self-regulatory resources can be depleted when engaging in tasks that are effortful but do not necessarily require suppression of affective reactions.

Resource depletion effects can also be observed between different purely cognitive tasks. Muraven et al. (1998) performed an experiment in which some subjects were instructed to not think of a white bear (Wegner et al., 1987), while controls were not told to suppress any particular thoughts. The dependent measure employed by Muraven

et al. (1998) was persistence on a set of unsolvable anagrams; subjects were instructed to spend as much time working on the anagrams as they wanted. Subjects who were instructed to suppress thoughts of a white bear spent less time (decreased persistence) attempting to solve the anagrams compared to controls who did not suppress thoughts in the first task.

Psychologists have also found that some social situations that require self-control can induce resource depletion effects. Richeson and Shelton (2003) examined the impact of interracial interaction on subsequent self-control ability, hypothesizing that more biased individuals would exert more self-control when interacting with a different-race confederate compared with a same-race confederate. White subjects first performed an implicit association task (IAT; Greenwald, McGhee, & Schwartz, 1998), enabling measurement of racial prejudice, and then they interacted with a confederate (Black or White) on a racially sensitive topic. After this phase, all subjects performed the Stroop task color naming task (MacLeod, 1991), in which subjects must suppress the automatic tendency to read words and instead must state the color the words are printed in. The analysis revealed two effects: first, subjects who interacted with White confederates showed no depletion effect; second, for subjects who interacted with a Black confederate, the degree of depletion observed on the Stroop task was predicted by the degree of bias measured with the IAT. In addition, Vohs, Baumeister, and Ciarocco (2005) performed several experiments examining the impact of effortful self-presentation on self-control. They found that effortful self-presentation (e.g., presenting oneself boastfully to friends) impaired subsequent performance on a difficult cognitive task. They also found that

initial exertion of self-control on a difficult cognitive task (i.e., the Stroop task) elicited poorer subsequent control over self-disclosure intimacy as measured by the Relationship Closeness Induction Task (Sedikides, Campbell, Reeder, & Elliott, 1998).

Taken as a whole, the aforementioned work demonstrates that the resource depletion effect is not confined to cognitive, affective, or social domains. Exerting self-control on a visual attention task impairs subsequent efficacy on both cognitive tasks (e.g., GRE problems; Schmeichel, Vohs, & Baumeister, 2003) and social overtures (e.g., impression management; Vohs et al., 2005). In addition, this work demonstrates that the same tasks used to induce depletion in one experiment (e.g., a sad video clip) are also valid dependent measures in other experiments (Schmeichel, 2007).

Extensions built upon the aforementioned studies have uncovered moderators of the depletion effect, possibly hinting at some of the mechanisms involved. Gailliot et al. (2007) found in several experiments that administration of glucose (sweet lemonade) attenuated depletion effects. Although the findings in Gailliot et al. (2007) have come into question (Kurzban, 2010), subsequent work has found that simply tasting but not ingesting glucose has similar restorative effects (e.g., Molden et al., 2012). Longitudinal studies in which subjects are asked to practice self-control exercises over the span of a few weeks, such as managing finances or study habits, have found that depletion effects can become attenuated with practice (Oaten & Cheng, 2006; 2007). These findings are used to support the dominant 'muscle' model of self-control exertion that accounts for depletion effects (Baumeister, Vohs, & Tice, 2007). Accordingly, not unlike skeletal muscles, self-control resources can be exhausted in the short term and also strengthened

as a result of long-term practice. Although other models have been proposed to account for observed depletion effects (e.g., Inzlicht & Schmeichel, 2012), compelling evidence to support them has yet to emerge.

Meta-Analyses and Failed Replications

In order to summarize the depletion studies published at the time, Hagger, Wood, Stiff, and Chatzisarantis (2010) conducted a meta-analysis to estimate the overall resource depletion effect size and to explore related phenomena (e.g., replenishment) associated with the muscle model. Published ego depletion studies up to that point in time (1998 - 2009) have shown that a wide variety of novel manipulations can produce robust effects. Hagger et al. (2010) estimated the overall depletion effect size as $d = 0.62$, 95% CI [0.57, 0.67], corresponding to a medium effect size according to guidelines from Cohen (1992).

However, in the following years well-powered studies employing commonly used manipulations and dependent variables have failed to replicate the robust effects described by Hagger et al. (2010), raising questions as to the accuracy of Hagger et al.'s estimate of effect size. Carter and McCullough (2013b) attempted to replicate depletion effects with a large ($N = 235$) sample and found significant *enhancement* in working memory performance for depleted subjects, a result completely at odds with previous studies. Xu et al. (2014) conducted four separate studies using tasks associated with the strongest effect sizes as described by Hagger et al. (2010). Across four studies, no significant results were obtained; in three of the four studies, effects again trended in the

opposite direction of expected depletion effects (enhanced performance in depleted subjects).

The failed replications mentioned above are completely inconsistent with the experimental effects described in the ego depletion literature and with the Hagger et al. (2010) meta-analysis; clearly, something is amiss. Carter and McCullough (2013a) proposed that publication bias inflated the effect size estimates from Hagger et al. (2010), and pointed out several flaws in the older meta-analytic techniques. A subsequent meta-analysis of ego depletion studies by Carter and McCullough (2014) employed enhanced methods to correct for the influence of publication bias, and concluded that ego depletion effects are not distinguishable from zero. In summary, there appears to be a great deal of uncertainty as to the true effect size associated with resource depletion effects.

Chapter Two: Behavioral Experiment

The aim of the current experiment was to obtain an accurate empirical estimate of the resource depletion effect size using a between-subjects design commonly seen in previous work. In the current experiment, subjects in the depletion condition first performed a video task shown by previous work to drain self-regulatory resources (e.g., Schmeichel et al., 2003), while controls performed an easier version of the same task. Following the video task, all subjects then performed the Stroop task, which served as the dependent measure. As the intent was to maximize sample size for a precise estimate of effect size, no a priori power analysis was conducted. Estimates of effect size are reported in addition to conventional significance tests, and a bootstrapping procedure was used to estimate 95% confidence intervals for effect size estimates.

Method

Subjects

A total of 273 subjects (76 male, 197 female) participated in the Behavioral Experiment. Subjects were recruited from the University of Minnesota Research Experience Program pool and compensated with partial class credit. Subjects were 18 years of age or older, native English speakers, and had normal or corrected-to-normal color vision. A total of 16 subjects were excluded from analysis (7 non-native English speakers, 2 failures to perform the Stroop task above chance, 6 instances of equipment failure, and 1 subject who performed a previous version of the experiment), leaving a total of 251 subjects for analysis. Subjects were randomly assigned to either the attention

manipulation (depletion, $n = 129$) or control ($n = 122$) groups in a 2-cell between-subjects design.

Materials

Video task. In the video task (Gilbert, Krull, & Pelham, 1988; Schmeichel, Vohs, & Baumeister, 2003), subjects were instructed to watch a brief (4 minute) silent video of an employment interview. They were told that they would later make judgments about the interviewee's personality based on their body language (however, no such measure was actually taken). Controls were instructed to simply watch the video, while subjects in the attentional manipulation (depletion) condition were given the additional instruction to avoid viewing the words presented on the bottom of the screen and to look away if they happened to look at one of the words.

During each video, a total of 24 common English words (see Appendix A) were presented in series on the bottom of the screen for 10 seconds each. Four wordlists were generated such that word frequency (Francis & Kucera, 1982) was roughly equal between lists (overall mean word frequency of 80.24 per million). Eight separate videos were constructed from the four wordlists (each with 24 words) and two source video clips (one male and one female interviewee), with wordlist and interviewee gender randomly assigned to subjects. Multiple videos were used to rule out the possibility that any observed effects were idiosyncratic to a particular video. A representative screen capture of the videos is shown in Figure 1.

Stroop task. Most of the parameters of the Stroop task in this experiment were drawn from those used by Inzlicht and Gutsell (2007). Stimuli consisted of two color

words (green or purple), and were presented in a font color that was either congruent (e.g., ‘green’ in green-colored letters) or incongruent (e.g., ‘green’ in purple-colored letters) with the word, giving a total of four possible stimuli. Individual letters that composed each word subtended approximately 0.5 degrees of visual angle.

Figure 1

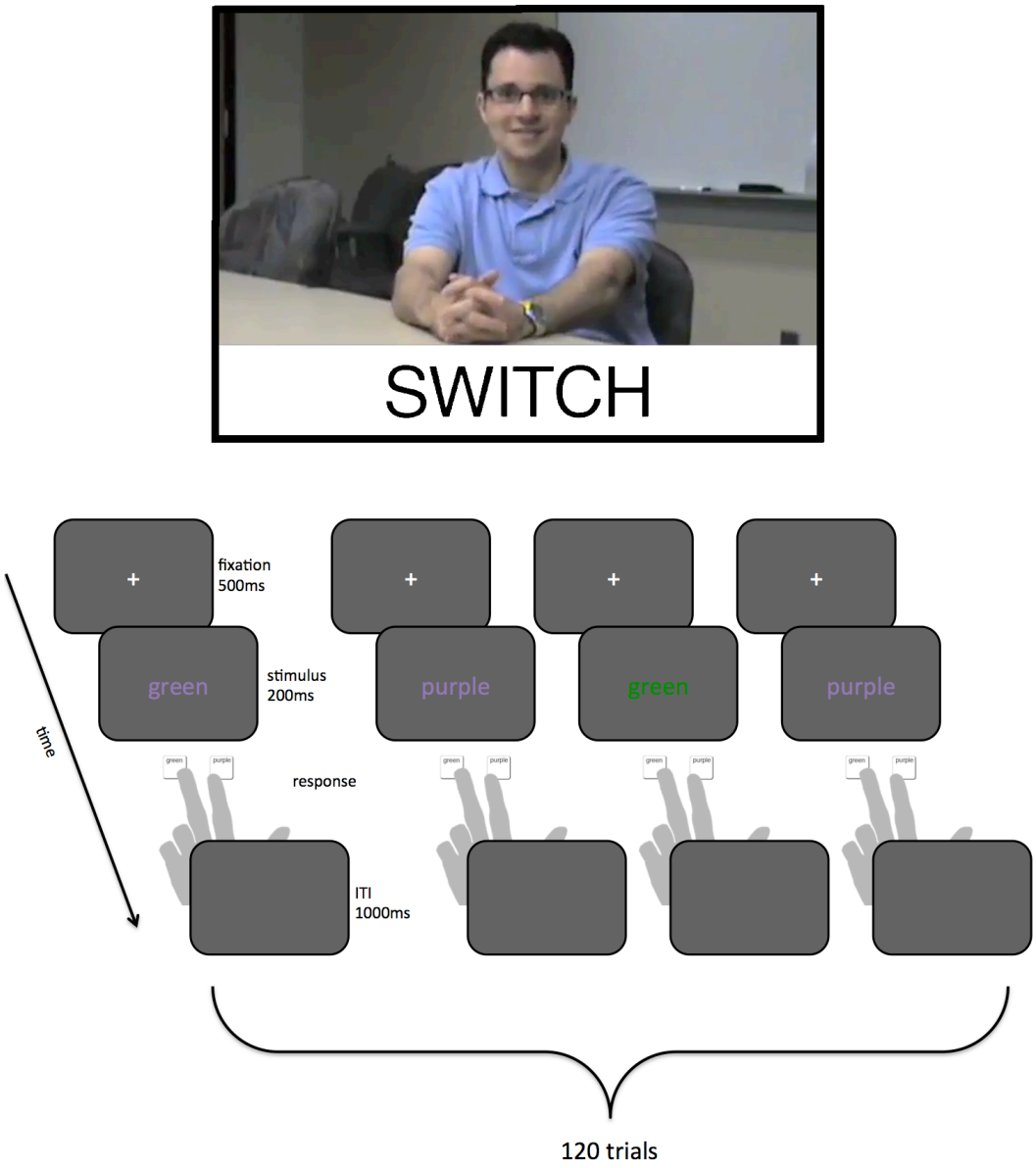

Figure 1. Sample frame from the attentional control video (top), and schematic showing Stroop task parameters (bottom).

Procedure

Upon arriving at the lab, subjects were randomly assigned to either the attentional manipulation (depletion) or the control condition; subjects within each group were randomly assigned to view one of eight possible videos (described above). Subjects completed the experiment individually with an experimenter in the testing room. After providing informed consent and basic demographic information, the experimenter handed the subject a sheet with printed instructions for the video and Stroop tasks (see Appendix B). The subject was instructed to read the sheet and return it to the experimenter when finished reading, at which point the experimenter asked the subject to verbally describe both tasks to ensure that the subject understood the instructions. Instructions for both tasks were presented at the beginning of the experiment to minimize the length of time between the end of the video task and the start of the Stroop task; previous resource depletion experiments have found that simple passage of time can attenuate depletion effects (e.g., Tyler & Burns, 2008). Instruction sheets for subjects in the depletion condition directed subjects to avoid looking at words during the video task, and if they did, to refocus their attention on the interviewee; all other instructions were identical between conditions (see Appendix B).

At this point, subjects were asked to place their chins in a chinrest. Subjects then performed 4 Stroop practice trials and watched a brief (approximately 15 s) portion of the attentional control video as practice. After the practice tasks, subjects then watched the attentional control video, with attentional control (depletion) subjects reminded to not look at the words on the bottom of the screen.

Once the video ended, subjects performed the Stroop task. The task consisted of 120 individual trials, 70% of which were incongruent and 30% of which were congruent. Trials were presented in a pseudorandom fashion, such that no single trial type (congruent or incongruent) or word color was presented for more than 4 consecutive trials. Subjects were instructed to respond using a button box to the color the word was printed in, and not the meaning of the word.

Each trial began with a fixation cross displayed for 500 ms. The Stroop stimulus word was then presented for 200 ms, followed by a blank screen at which time subjects would respond to the trial. After each response, no feedback was given, and the display remained blank for 1500 ms after which the fixation cross for the next trial was displayed. During the Stroop task, subjects kept their chins in a chinrest placed 50 cm from the computer monitor. A visual representation of the trial structure is shown in Figure 1.

Results

The Stroop task served as the dependent measure in this experiment. Per-subject interference scores were obtained for response times and error rates by subtracting each subject's mean response time or error rate to congruent trials from the mean for incongruent trials. For response times, trials with incorrect responses were removed, and response times greater than 2.5 *SD* from the overall per-subject mean were replaced with the 2.5 *SD* threshold value (truncated). Interference scores calculated in this fashion are thought to reflect the difficulty of exerting cognitive control on incongruent trials, and are the most common measures of Stroop performance (MacLeod, 1991).

Figure 2

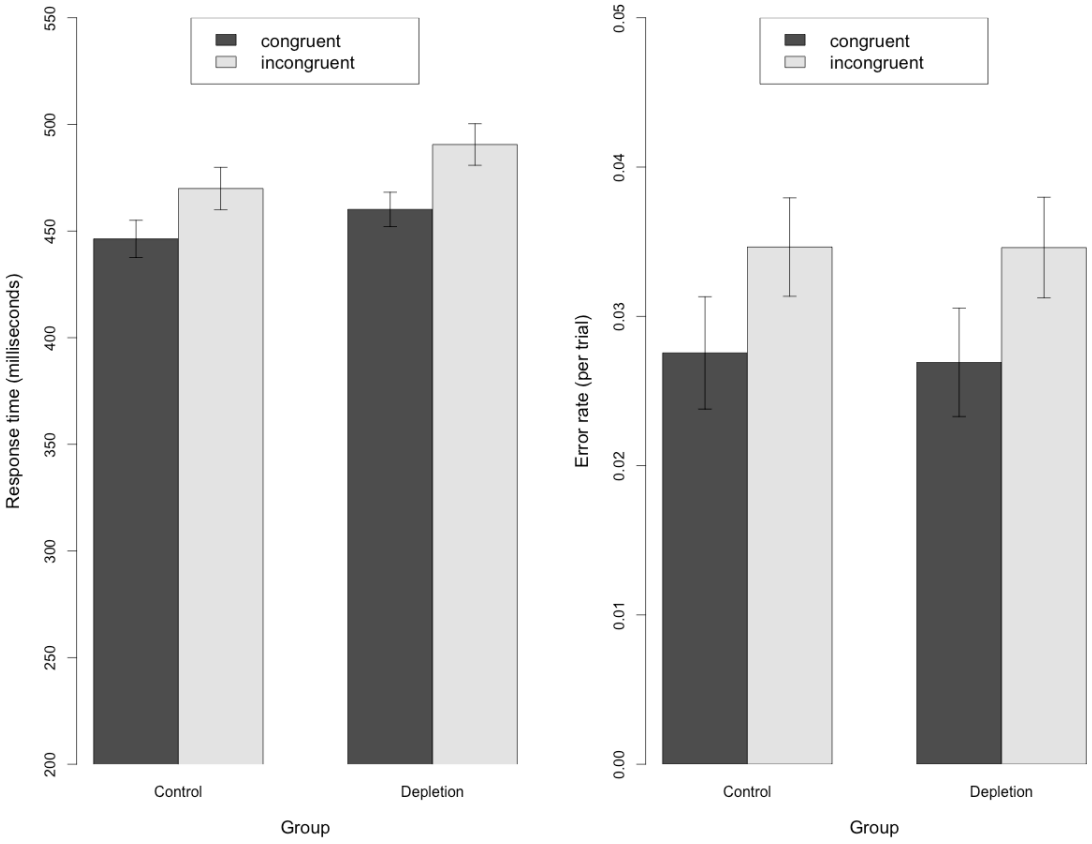

Figure 2. Bar plots showing mean response times (left) and per-trial error rates (right) by group and trial type in the Stroop task.

Table 1

Mean Response Times in the Stroop Task, by Group (Standard Deviation in Parentheses)

	N	Interference Effect	Congruent	Incongruent
Depletion	129	30.43 (34.65)	460.14 (91.59)	490.57 (110.79)
Control	122	23.61 (34.56)	446.33 (96.55)	469.94 (109.48)
Total	251	27.12 (34.10)	453.42 (94.10)	480.55 (110.42)

Table 2

Error rates in the Stroop Task, by Group (Standard Deviation in Parentheses)

	N	Interference Effect	Congruent	Incongruent
Depletion	129	.0076 (.0450)	.0269 (.0412)	.0346 (.0382)
Control	122	.0070 (.0436)	.0275 (.0415)	.0346 (.0364)
Total	251	.0073 (.0442)	.0272 (.0413)	.0346 (.0373)

Hypothesis Tests

Collapsing across depletion and control groups, response times to congruent trials ($M = 453.42$, $SD = 94.10$) were significantly ($t(250) = 12.381$, $p < .001$) faster compared to responses for incongruent trials ($M = 480.55$, $SD = 110.42$). Similar effects were found for error rates; subjects were significantly more likely to commit errors on incongruent ($M = .0346$, $SD = .0373$) trials compared to congruent ($M = .0272$, $SD = .0413$) trials ($t(250) = 2.646$, $p = .004$).

A one-tailed between-subjects *t* test was used to test the hypothesis that subjects in the depletion condition performed more poorly on the Stroop task compared to controls. Response time interference scores in the depletion group ($M = 30.43$, $SD = 34.65$) were greater than those in the control group ($M = 23.61$, $SD = 34.56$), but this difference did not reach significance, $t(249) = 1.56$, $p = .0599$. A second between-subjects *t* test on interference for error rates was also conducted. Subjects in the depletion group ($M = .0076$, $SD = .0450$) did not exhibit increased interference compared to control subjects ($M = .0070$, $SD = .0436$; $t(249) = -0.107$, *ns*). Separate exploratory ANOVAs conducted on each dependent measure with experimental condition, wordlist, and interviewee gender as factors did not yield any significant main effects or interactions.

Figure 3

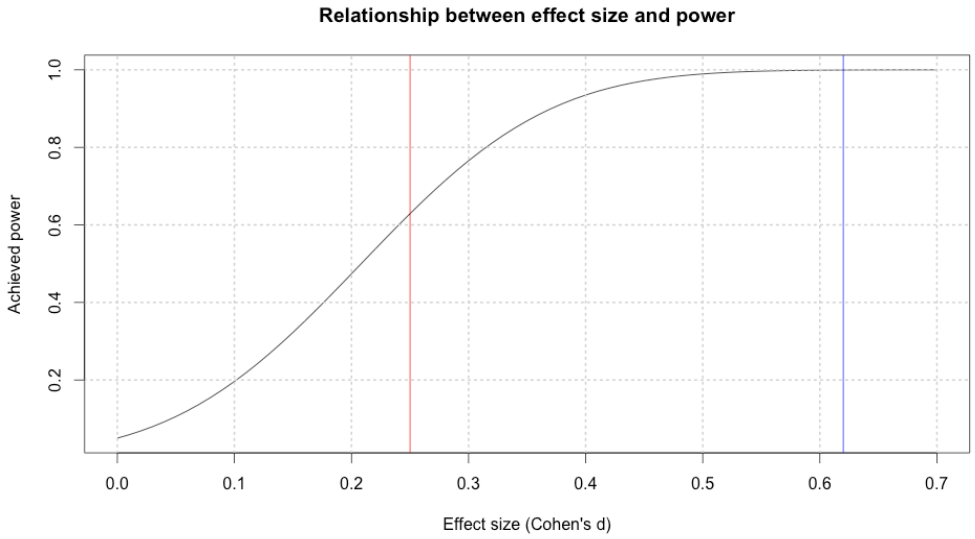

Figure 3. Plot illustrating the relationship between achieved power and effect size for a one-tailed *t*-test in the Behavioral Experiment. The red and blue lines show effect size estimates from Carter and McCullough (2014) and Hagger et al. (2010) respectively. Figure produced using G*Power (Faul et al., 2007).

Post-hoc power analyses were conducted to determine the achieved power to reject the null hypothesis. Assuming an effect size of $d = 0.62$ (as suggested by Hagger et al., 2010), the Behavioral Experiment achieved a power level of 99.94% (one-tailed) to reject the null hypothesis. If, however, the true effect size is assumed to be $d = 0.25$ (as suggested by the PEESE analysis in Carter & McCullough, 2014), the power level achieved was 62.90% (one-tailed). A plot of the relationship between effect size and achieved power (given the sample size from the Behavioral Experiment) is shown in Figure 3.

Bootstrapped Sampling Distribution

Estimates of between-group effect size and associated confidence intervals for Stroop interference were calculated using the bootES R package (Kirby & Gerlanc, 2013). Generally, bootstrapping procedures estimate confidence intervals by repeatedly drawing random samples (resamples) with replacement from the set of original data points, and calculating an estimate of effect size for each resample. The distribution formed by effect size estimates from many resamples then forms the sampling distribution. The bias-corrected-and-accelerated bootstrapping procedure (Efron, 1987) improves upon simple bootstrapping methods by compensating for biased estimators of population parameters, as well as nonindependence of standard error estimates with regard to population parameters. The BCa method has been shown to accurately estimate the sampling distribution of Cohen's *d*, and is robust to violations of normality (Kelley, 2005).

A total of 100,000 resamples were obtained from the per-subject Stroop response time interference scores. The between-subjects effect size (calculated from the group means and deviations) was $d = 0.197$. This is considered a 'small' effect according to guidelines proposed by Cohen (1992). The bootES package calculated the 95% confidence interval as [-0.050, 0.484]. Critically, the bounds of the bootstrapped 95% CI include 0, and exclude the $d = 0.62$ (95% CI: 0.57, 0.67) estimate by Hagger et al. (2010).

Figure 4

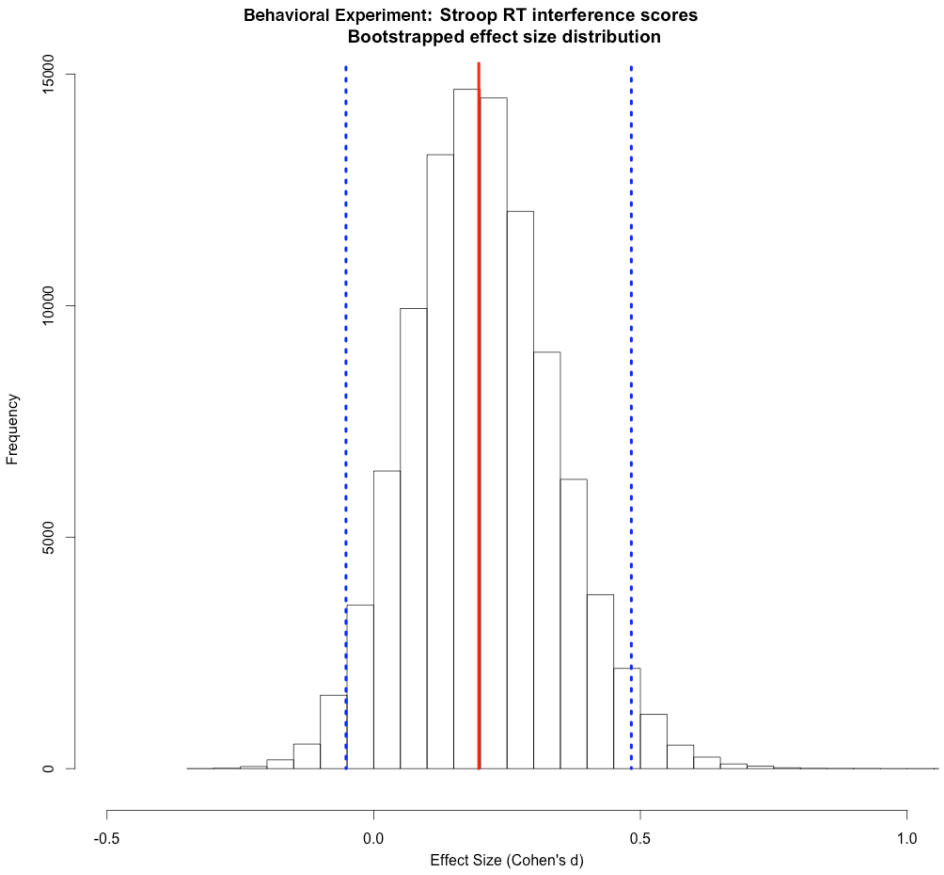

Figure 4. Histogram of the bootstrapped sampling distribution of Stroop RT interference scores. The red vertical line shows the observed effect size in the Behavioral Experiment; blue dashed lines show the 95% CI obtained with the bootstrapping procedure.

Chapter Three: Meta-Analysis and Publication Bias

Touted in part as a way to reconcile seemingly incompatible findings obtained from different individual studies, meta-analysis has been widely used since its introduction in the late 1970s (Glass, 1976). However, the validity of the results produced by standard meta-analysis is threatened by publication bias, which favors the publication of statistically significant results and may bias meta-analytic effect size estimates upward. Recent work has shown that the most common method used to correct for the effects of publication bias in meta-analysis (trim and fill; Duval & Tweedie, 2000a; 2000b) performs poorly in cases in which underpowered studies test relatively small effects. A new meta-analytic method, *p*-curve, was developed by Simonsohn, Nelson, and Simmons (2014) to produce effect size estimates unaffected by publication bias. Their simulations demonstrated that *p*-curve effect size estimates are unaffected by publication bias, but simulation performance degrades when questionable research practices are introduced.

Evaluating Multiple Studies before 1976: Ad-Hoc Methods, Narrative Reviews

The accumulation of early empirical studies in psychological research prompted researchers to write literature reviews that examined and summarized multiple studies on a given topic, a practice that continues to this day. When reviewing and synthesizing the available literature or studies on a given topic, some authors included ad-hoc or informal statistical analyses of the studies in question (e.g., Underwood, 1957). One method involved tallying the number of studies in a box-score fashion that obtained significant results in favor of, or opposed to, a specific hypothesis (Light & Smith, 1971). Fisher (1932) described a method to combine *p*-values obtained from a number of independent

tests that appeared to show marginal results but failed to reach significance. This method was further developed by Mosteller and Bush (1954), but neither their revision nor the original method saw widespread use in psychology.

Explosion of Research

Early methods for quantitative evaluation of results from multiple studies (some of which are described above) were employed on an ad-hoc basis and were not recognized as a unique class of statistical procedures until the mid-1970s. In the years following the conclusion of World War II, there was a virtual explosion in the quantity of published research. For example, the number of journals in psychology increased from 91 in 1951 to a total of 1195 journals by 1992 (Olkin, 1995). The number of published studies on a given topic grew so large as to preclude a reviewer from synthesizing those studies into a cohesive whole, even if time were available for the actual reading. Furthermore, synthesis of published research became difficult when multiple studies of the same phenomena yielded different and opposing results. When this occurs across hundreds of studies on a topic of interest, it is questionable as to whether or not an individual author can be relied upon to ingest, recall, and evaluate the evidence at hand accurately and without bias (Shadish & Lecy, 2014).

Advent of Modern Meta-Analysis

By the 1970s, researchers in psychology and other fields struggled to make sense of and integrate a growing and often discordant literature. Government policymakers and the public sought answers to important social problems (e.g., efficacy of educational interventions), and were frustrated with reports of conflicting results (Hunter & Schmidt,

2014). Gene Glass, in his 1976 presidential address to the American Educational Research Association, coined the term 'meta-analysis', defining it as the analysis of summary statistics (and not raw, per-subject data) drawn from many separate studies (Glass, 1976). The general framework of meta-analysis proposed by Glass (1976) bore some similarity to the methods used for per-subject data in empirical studies, which may have contributed to its popularity. At the time of Glass's (1976) address, there was no consensus as to what might be the best or most statistically rigorous way to perform meta-analysis. Gene Glass's collaboration with Robert Rosenthal and Frank Schmidt (who at the time of Glass's address were independently developing ways to integrate data from multiple studies; Shadish & Lecy, 2014) ultimately converged on the meta-analytic methodology in common use today. Briefly, this method involves calculation of a mean effect size that is weighted by a per-study measure of variability (e.g., sample size). An estimate of the error between studies is also calculated, making it possible to obtain a confidence interval around the weighted mean effect size (Hunter & Schmidt, 2014; Hedges & Vevea, 1998).

Ability to Test Effects of Moderators

In addition to a statistically rigorous way to combine measures of effect size, meta-analysis lets researchers examine and test for the effects of moderators on a set of findings. For example, imagine that some studies showed that differences between high-SES and low-SES elementary students on a test of verbal ability were attenuated when students were given free breakfast, while other studies failed to show the breakfast-related enhancement. Just as how an overall or omnibus meta-analysis produces an

overall effect size estimate and confidence interval, separate meta-analyses can be conducted for subsets of studies. The resulting estimates and confidence intervals can then be directly compared to evaluate whether or not (for example) the breakfast effects are real, or attributable to sampling error. Effects of moderators along a continuous scale can be evaluated using various forms of regression. This feature is often the primary focus of modern published meta-analyses, as it can establish or clarify the relationship between two separate phenomena or treatments.

What are meta-analyses used for, aside from examining the effects of moderators? Accurate estimates of overall effect size have much practical utility when planning a study and performing a priori power analyses. Outside of the practical aspect, accurate effect size estimates have little bearing on theoretical development or falsification. Most theories in psychology predict relationships between variables or changes in behavior only in a directional fashion, and largely do not make point predictions as to the magnitude of any given observation or relationship between variables (Meehl, 1967). Any statistically significant result in the right direction supports theory, regardless of the magnitude. An observed effect size of $d = 0.3$ supports a proposed theory just as well as an effect size of $d = 0.8$. For this reason, the accuracy of an overall effect size estimate has not been a major cause for concern provided that the effect can be differentiated from zero.

Publication Bias

Publication bias refers to the phenomenon in which researchers, editors, and reviewers decide which papers to publish based mainly on the positive results they

contain. Although the probability of publication may be dependent on a variety of factors (e.g., effect size, author prestige), statistical significance appears to be a de facto requirement for publication (Sterling, Rosenbaum, & Weinkam, 1995). The apparent difficulty of interpreting nonsignificant results in addition to the limited space available in journals may be contributing factors to publication bias. Unsurprisingly, publication bias introduces problems for the meta-analyst.

Effects of Publication Bias on Meta-Analysis

Unpublished results are, by definition, more difficult to uncover and obtain compared with published ones. When nonsignificant results are not published, a meta-analyst is not sampling from the population of all experiments conducted on a given phenomenon; instead, sampling is biased by publication as a proxy for accessibility. Even when unpublished experiments are obtainable, the time and effort required to obtain them can easily be imagined to exceed that required for published work.

A brief example here may help illustrate the nature and magnitude of the problem. Figure 5 shows sets of between-subjects *t* tests that were obtained using two different hypothetical experimental effects with $d = 0.8$ and $d = 0.3$. Despite the fact that these two hypothetical effects differ greatly in magnitude, the mean effect size for significant tests is roughly the same for both. This shows that when only significant results are published, a true effect size of (e.g., $d = 0.3$) can mimic the appearance, in terms of average published effect sizes, of a much larger effect.

Figure 5

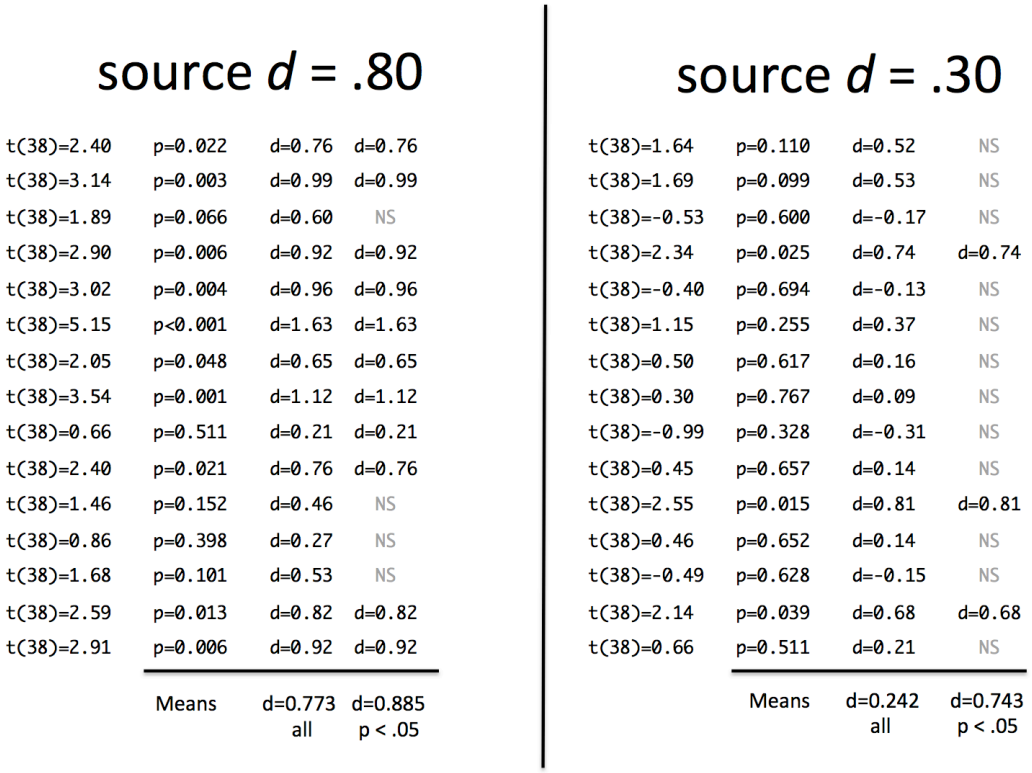

Figure 5. Effects of publication bias on mean effect size.

Detection of Publication Bias and Correction for Influence

The almost total absence of nonsignificant findings in published psychological reports (Sterling, 1959; Sterling et al., 1995) suggests that publication bias is a widespread phenomenon. Early methods to address potential effects of publication bias in meta-analysis include Rosenthal's fail-safe *N* (Rosenthal, 1979), which calculated the number of studies with null (average) effect sizes that would need to be added to the meta-analysis for the overall significance test for the effect to fall above $p = .05$. A small

failsafe *N* suggests that few file-drawered null results need exist to call the meta-analytic evidence for the finding into question; determination of how many studies qualifies as 'small' is left to the reader. A number of other methods were proposed to correct for publication bias but have been rarely employed (e.g., Iyengar & Greenhouse, 1988), likely due to the complex models involved and questionable assumptions as to the mechanics of publication bias (Dear & Dobson, 1997; Harris & DuMouchel, 1997). The most widely used methods for detection of publication bias and correction are based on the funnel plot (Light & Pillemer, 1984).

Figure 6

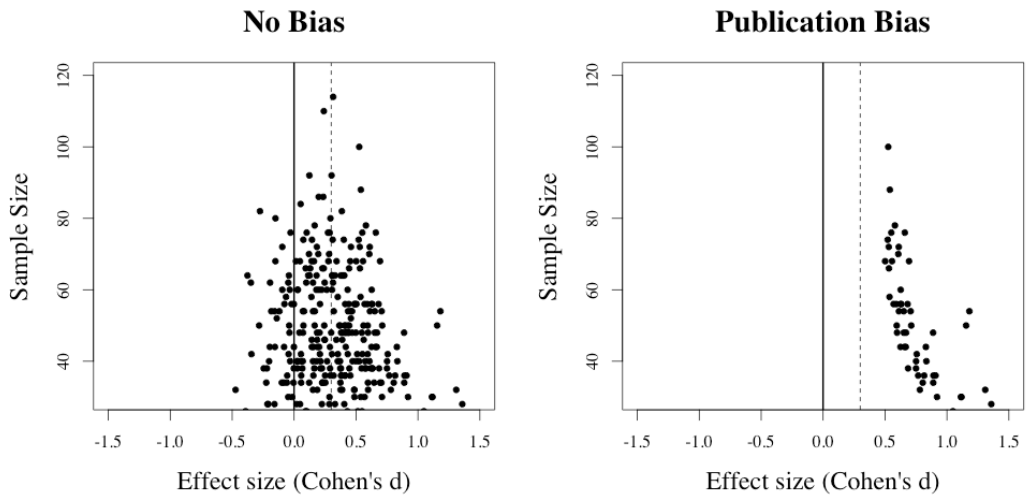

Figure 6. Funnel plots of studies testing an effect size of $d = 0.3$. Plot on the left shows no bias (all studies are plotted, regardless of significance), while the plot on the right shows only studies that have reached conventional significance levels.

The funnel plot shows the relationship between the distribution of per-study effect sizes in a meta-analysis, the precision of each study, and the role of statistical

significance. Each study in a meta-analysis is represented as a single point. Studies are arranged along the x-axis by effect size (here, using Cohen's *d*), and along the y-axis by some measure of precision, such as standard error or sample size. The intuitive appeal of the funnel plot comes from the relationship between precision and effect size: as precision increases along the y-axis, observed effect sizes decrease in variability, forming an inverted V or funnel shape.

If the availability of studies is not dependent on effect size the funnel plot will exhibit symmetry as in Figure 6 (left panel). When publication bias removes studies with small effects or nonsignificant results, as in Figure 6 (right panel), the lack of symmetry is plainly visible. Of course, simple visual inspection of a funnel plot is subjective and prone to bias (Villar, Piaggio, Conneli, & Donner, 1997).

Formal statistical methods have been proposed to detect funnel plot asymmetry. Begg and Mazumdar (1994) found that publication bias can result in a significant correlation between per-study measures of effect size and variance, and that presence of such a correlation is suggestive of publication bias. A similar method using linear regression was shown by Egger, Smith, Schneider, and Minder (1997) to differentiate between biased and unbiased sets of studies (confirmed by concordant or discordant effect sizes from very large clinical trials).

In addition to detecting publication bias, funnel plot-based methods can also be used to correct inflated effect size estimates with the trim and fill method (Duval & Tweedie, 2000a, 2000b). The trim and fill method is an iterative process that entails first estimating the number of unpublished (unavailable) studies, removing (trimming) that

number of studies with the largest effect sizes from the plot until a symmetric plot is obtained, and finally adding studies (filling) on either side of the mean effect size of the trimmed (symmetric) plot. Compared to other, more complex corrections for publication bias (e.g., Given, Smith, & Tweedie, 1997), trim and fill does not rest upon questionable assumptions as to the mechanics of publication bias, nor does it require complex computational procedures. The tractability of trim and fill has apparently made this method the most popular way to correct for publication bias in meta-analysis.

How effective is trim and fill at adjusting effect size estimates to account for publication bias? Simulations performed by Simonsohn, Nelson, and Simmons (2014) clearly show that trim and fill does a remarkably poor job of correcting inflated effect size estimates. In cases in which the true effect size is small (i.e., only underpowered studies are performed), and only significant results are available, the per-study effect sizes must be very large relative to the true effect size in order to achieve statistical significance (see Figure 5). The trim and fill method effect-size estimate is obtained by computing the mean effect size for a subset of available studies (specifically, those studies that are not trimmed). If the true effect size is below the smallest effect size observed in a set of studies, trim and fill cannot trim its way to a mean effect size that is outside the range of per-study effect sizes.

For example, imagine trim and fill is used on a set of studies with effect sizes d = (0.40, 0.45, 0.38, 0.51, 0.44). Regardless of which studies are trimmed from that set, the mean of the remaining studies can never be less than $d = 0.38$. If the set of studies

described above tested a true effect size of $d = 0.30$, for example, trim and fill would never be able to recover the underlying effect size.

p-Curve: Correcting for Publication Bias with Significant Studies Alone

Recent work by Simonsohn, Nelson, and Simmons (2014) has focused on the analysis of the distribution of significant *p*-values across a set of studies. Due to publication bias, information about studies with $p > .05$ is difficult to obtain. Recognizing this, Simonsohn et al. have developed analysis methods that utilize significant studies alone, thereby sidestepping the issue of publication bias and study availability. Initial efforts at examining *p*-value distributions focused on assessment of the evidential value of a set of studies (Simonsohn et al., 2014). Subsequent work has described a meta-analytic procedure, *p*-curve, which estimates effect size in a way that does not directly take into account per-study effect sizes. Monte Carlo simulations have shown that *p*-curve estimates of effect size are remarkably accurate under conditions of publication bias, while other correction methods (e.g., trim and fill) fare much more poorly (Simonsohn et al., 2014).

How does *p*-curve produce unbiased estimates of effect size from published studies alone? With *p*-curve, effect sizes are estimated by relying on a fundamental property of null hypothesis significance testing: the distribution of *p*-values is uniform when the null hypothesis is true. A large set of studies testing null hypotheses (i.e., $d = 0.0$) will contain the same number of *p*-values between .70 to .75 as compared to between .00 and .05 or between .20 to .25. Publication bias limits the studies we can 'see' to a narrow window in which *p*-values are between .05 and .00; however, even within that

narrow range, *p*-values are still uniformly distributed under the null hypothesis. If the null hypothesis is not true, the distribution of *p*-values will be skewed to the right (i.e., many *p*-values between .00 and .01). A set of studies with a uniform *p*-curve strongly suggests that those studies are a collection of Type I errors.

Figure 7

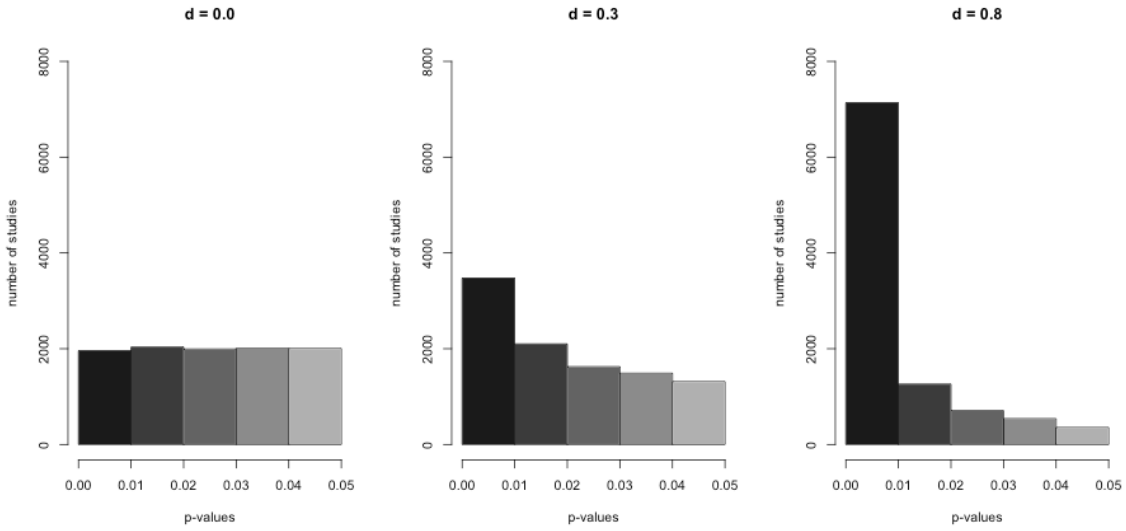

Figure 7. Each panel shows 10,000 significant p-values with underlying effect sizes of $d = 0.0$ (left), $d = 0.3$ (center), and $d = 0.8$ (right). Per-cell sample sizes were drawn from a uniform distribution ranging from N=10 to N=50.

Within each study, as power to obtain a significant result increases, the distribution of significant *p*-values shifts, increasing the probability of observing low *p*-values (e.g., .01) relative to higher *p*-values (e.g., .045). This produces a nonuniform, skewed *p*-curve, which (ignoring sampling error) is inconsistent with a collection of Type I errors, but consistent with the presence of true non-zero effects. In a nutshell, *p*-curve works by generating simulated *p*-curves associated with a range of underlying effect

sizes, and finding the closest match to the observed *p*-curve; the effect size used to generate the best-fitting simulated *p*-curve is the estimate of effect size.

The vast majority of null hypothesis significance tests are conducted using a null hypothesis of no difference between groups or an effect size of zero. It is possible, but uncommon, to conduct a significance test in which the null hypothesis is that a difference of a given magnitude between groups exists. If a set of independent significance tests were conducted against a null hypothesis effect size equivalent to the population effect size, the *p*-value distribution would be uniform (ignoring any variation due to sampling error). *p*-Curve operates in a generally analogous fashion.

With *p*-curve, effect size is estimated by taking the observed *t*-statistics and degrees of freedom from a set of studies and calculating *pp*-values (*p*-value of the *p*-value), representing the probability of observing a significant *t*-statistic at least as large given a candidate effect size. If the underlying effect size is assumed to be zero, a uniform distribution of significant *p*-values corresponds to a uniform distribution of *pp*-values from 0 to 1. If large effect sizes are assumed, observing *t*-statistics larger than those associated with *p*-values of .04 or .05 is very likely; in this case relatively few *p*-values between .04 and .05 should be observed compared to *p*-values from .00 to .01.

For example, imagine a study reports a *t*-statistic of 2.40 with 38 degrees of freedom, with an associated *p*-value of .01. The *p*-value reported tests a null hypothesis, and represents the probability of encountering a *t*-value at least as extreme as the one observed assuming the null is true. To obtain the *pp*-value for a candidate effect size of $d = 0.4$, first we calculate the power, or the probability of obtaining a significant ($p < .05$)

result and obtain 0.23. In the same fashion we calculate the probability of obtaining a *p*-value equal or smaller than the observed *p*-value, again assuming a candidate effect size of $d = 0.4$, and obtain 0.14. The *pp*-value is obtained by dividing the probability of observing a smaller *p*-value over the power of the test to achieve significance, giving us a *pp*-value of 0.60. Assuming that the true underlying effect size is $d = 0.4$, the *pp*-value tells us that the probability of observing a *p*-value at least as small, *assuming a significant result has been obtained*, is 0.60. The *pp*-values obtained in this fashion allow direct comparison of *pp*-values from many studies with varying sample sizes. When the candidate effect size is equal to the true underlying effect size, the *pp*-value distribution becomes uniform (see Figure 8).

Figure 8

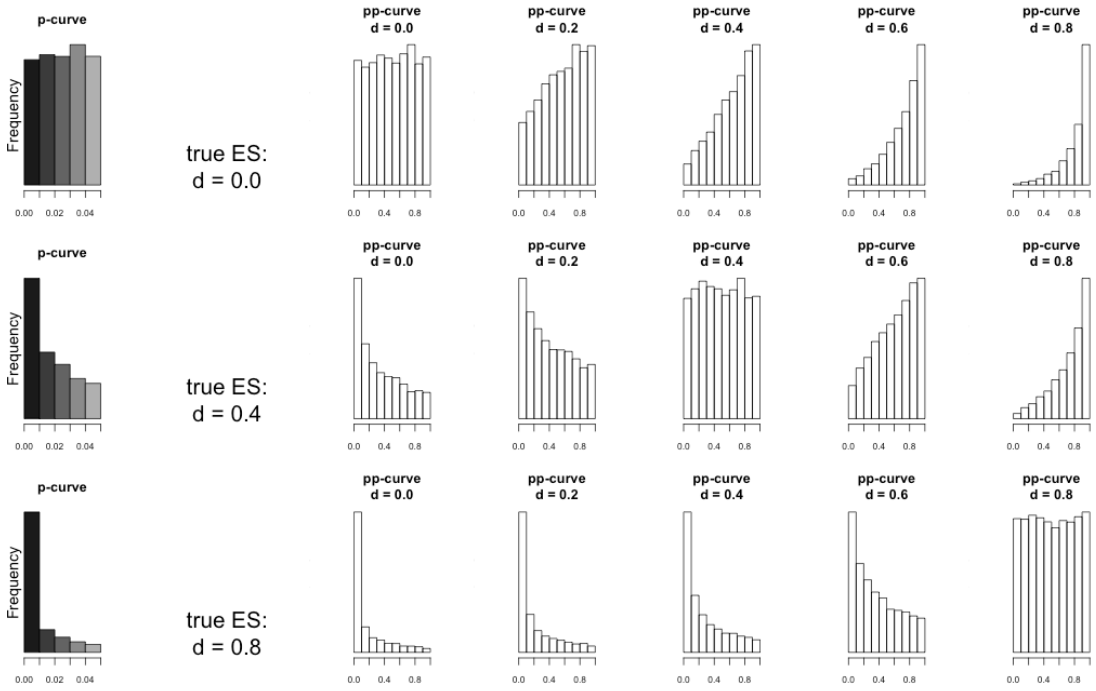

Figure 8. Illustration of *p*-curves and *pp*-curves at three different underlying effect sizes (*d* = 0.0, 0.4, 0.8); five candidate effect sizes are shown for each set of simulated studies. Note how the *pp*-curve is uniform when the true effect size is equal to the candidate (assumed) effect size.

The *p*-curve estimate of effect size is obtained by finding the candidate effect size that produces the most uniform distribution of *pp*-values. *Pp*-values are calculated at each possible effect size within a given range (e.g., *d* = -2.0 to 2.0), and a Kolmogorov-Smirnov test is used to quantify the degree to which the recalculated *pp*-value distribution is similar to a uniform distribution. The assumed effect size associated with the smallest Kolmogorov-Smirnov test statistic corresponds to the *p*-curve estimate of effect size (Simonsohn et al., 2014).

p-Curve: Assumptions and Limitations

The accuracy or validity of *p*-curve effect size estimates rests on the assumption that the distribution of *p*-values is uniform under the null hypothesis. At first glance this assumption might be a bit confusing – *p*-values are by definition uniformly distributed under the null due to the mechanics of null hypothesis significance testing. Problems arise, however, in cases in which comparisons are made between groups that are part of a larger experimental design that must reach significance in order to be published. In these cases, *p*-values under the null hypothesis cannot be assumed to be uniform (Simonsohn et al., 2014).

For example, imagine a researcher is interested in whether or not resource depletion effects (Baumeister et al., 1998) are attenuated or eliminated after exposure to cold air. To test this assumption, the researcher conducts a between-subjects experiment that independently manipulates cold air exposure (warm or cold) and task difficulty (untaxing or depleting), resulting in 4 unique experimental conditions. Subjects in two conditions perform difficult (vs. easy) self-regulatory tasks without exposure to cold air, while subjects in two other conditions are exposed to cold air before performing the dependent measure for the depletion task. The experiment 'works' if subjects exposed to cold air fail to exhibit a resource depletion effect, while controls do exhibit between-group differences in performance associated with resource depletion effects; a significant interaction effect between the two conditions would confirm this. If this significant interaction is not observed, results do not support the researcher's hypothesis on attenuating effects of cold air, and the 'failed' experiment is not publishable as a result.

In order for the interaction to be significant, the two groups not exposed to cold air must independently exhibit a resource depletion effect that is associated with a *p*-value *below* .05 – that is, the significant interaction requires a two-group comparison between controls to exceed the significance threshold. This causes the distribution of *p*-values under the null for the two-cell test to deviate from uniformity, thus invalidating *p*-curve's effect size estimates. This presents a substantial practical limitation on the use of *p*-curve to estimate effect size in cases in which most studies examine moderators of an effect using more complex experimental designs.

Chapter Four*: p*-Curve Meta-Analysis

In this analysis, two datasets were compiled and effect sizes were estimated using *p*-curve (Simonsohn et al., 2014). The first dataset (HWSC dataset) consists of all *p*-curvable studies used in the Hagger et al. (2010) meta-analysis. The second dataset (post-HWSC) includes *p*-curve amenable studies published after Hagger et al. (2010) obtained from reference database queries. In order to be consistent with previous analyses of resource depletion effects, the methods and search criteria used here mirror those used by Hagger et al. (2010) and Carter (2013) as closely as possible given the unique requirements of *p*-curve analysis.

Inclusion and Exclusion Criteria

All included experiments were required to meet several methodological and statistical criteria for inclusion. Articles were required to be published in a peer-reviewed journal and written in the English language. Methodologically, studies were required to test resource depletion hypotheses by manipulating exertion of self-regulatory resources between subjects. In particular, subjects in one group had to have performed a difficult task requiring self-control, while control subjects had to have performed an easier version of the same task; after the first task, all subjects then had to have performed a second self-control task that served as the dependent measure. Any tasks described by the authors as requiring (or measuring) self-control resources were deemed acceptable. Subjects recruited for each study were required to be from sources typically used in psychological research (i.e., undergraduate subject pools). Studies with unusual subject populations

(e.g., children or animals) were excluded, as self-regulatory capacity may operate in a different fashion in these populations (von Hippel & Henry, 2011).

Compared with traditional meta-analyses, additional constraints are imposed by the assumptions necessary for unbiased *p*-curve estimates of effect size. Most prominently, *p*-curve requires that all included studies reach statistical significance ($p \leq .05$); *p*-values in excess of .05 cannot be included. In addition, all included studies must use a 2-cell between subjects experimental design to ensure independence of *p*-values (as described in the preceding section). This constraint, which is not present in traditional meta-analyses, has the effect of excluding many published resource depletion experiments from the analysis; however, sufficient numbers of studies were found.

Literature Search

HWSC Dataset. As mentioned above, all studies that appeared in the 83 publications included in the Hagger et al. (2010) analysis were examined to determine if inclusion criteria were met. A total of 36 studies were found to meet inclusion criteria and were coded for analysis as described below.

Post-HWSC Dataset. In order to update the set of studies published following Hagger et al. (2010), abstract searches were conducted using several databases to find relevant articles. The search strategy and keywords used were based on the method employed by Carter (2013), except that searches for unpublished work (to correct for the effects of publication bias) were *not* conducted. A diagram of the study search and selection process is depicted in Figure 9. All searches were limited to articles published between April 1, 2009 (the cutoff date used by Hagger et al., 2010) and Dec. 31, 2013.

For all searches, the following search strings were entered (comprehensive details for all searches can be found in Appendix C): "self regulat*," "resource deplet*," "depleted resource*," "ego depletion," "limited resource*," and "regulatory resource*." These search terms were used to query the OVID, Web of Science, and EBSCOhost databases.

Results from the searches of all databases were combined and duplicate entries were removed, leaving a total of 6,324 publications. Abstracts and reference information were obtained for each publication and stored in a BibDesk (version 1.6.3; McCracken, 2015) database. Abstracts were examined to determine whether each publication was likely to contain studies testing resource depletion effects. For all publications identified as likely containing resource depletion experiments, each article was read in full to determine whether studies met the inclusion criteria described above. A total of 401 articles were identified as possibly containing relevant experiments; examination of each article for studies meeting inclusion criteria yielded a total of 39 studies. Those studies were coded for analysis as described below.

Figure 9

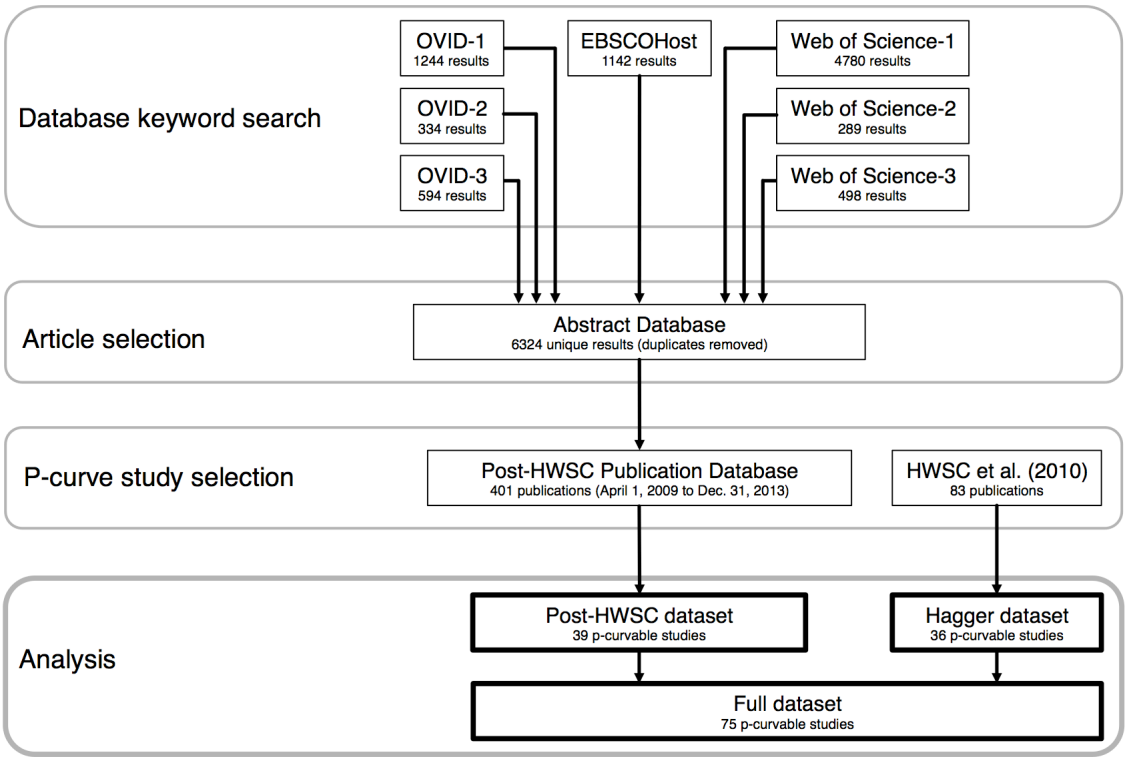

Figure 9. Diagram of study selection process.

Coding Procedures and *p*-Curve Disclosure Table

As recommended by Simonsohn et al. (2014), relevant information from each study (e.g., *t*-statistics, degrees of freedom) was coded into a *p*-curve disclosure table (see Appendix D for abbreviated version; the full version is included in the Online Supplement, available online at http://dx.doi.org/10.13020/D62S33). In cases where results were reported for multiple dependent variables (e.g., speed and accuracy on a Stroop task), results from the first statistical test to appear in the text were chosen. Every study was assigned a unique four-digit identifier when coded.

Statistical Methods

Traditional Meta-Analysis. Traditional meta-analytic procedures were modeled after those employed by Hagger et al. (2010) and implemented using the R metafor package (version 1.9-5; Viechtbauer, 2010). Exact sample sizes for each group were not coded in the *p*-curve disclosure table (often they did not appear in the original articles) and were assumed to be equal between groups. Effect size (Cohen's *d*) was calculated from the *t*-statistic and degrees of freedom as

$$d \approx \frac{2t}{\sqrt{df}}$$

A random-effects meta-analytic model was determined to be the most appropriate for this analysis. Resource depletion effects have been obtained using a variety of tasks, and the effect size associated with each task is not assumed to be identical. In fixed-effects meta-analysis, effect sizes are assumed to be identical, and variation in observed effect size between studies is attributed to sampling error. Random-effects meta-analytic methods, however, do not assume that all effect sizes across studies are identical (Hedges, 1992), which can be assumed to be the case with resource depletion tasks.

***p*-Curve Analysis**. Several R functions were written to implement the *p*-curve analysis, based on the R loss function published by Simonsohn et al. (2014). R code can be found in the Online Supplement located at (http://dx.doi.org/10.13020/D62S33). In order to obtain a confidence interval around the effect size estimate, a bootstrapping procedure similar to that used in the Behavioral Experiment was used. 100,000 samples with replacement were taken and *p*-curve estimates of effect size were calculated for each

subsample; the distribution formed by these estimates approximates the sampling distribution and was used to calculate the confidence interval.

Results

HWSC Dataset. The *p*-curve estimate of effect size for the HWSC dataset (all studies meeting criteria published before April 1, 2009) was $d = 0.18$, with the bootstrapped 95% confidence interval [-0.04, 0.49]. The *p*-curve plot for the HWSC dataset is skewed slightly to the right, suggesting the presence of a small effect (see Figure 10). For comparison, a random-effects meta-analysis conducted on the same dataset estimated the overall effect size as $d = 0.82$, 95% CI [0.75, 0.90].

Figure 10

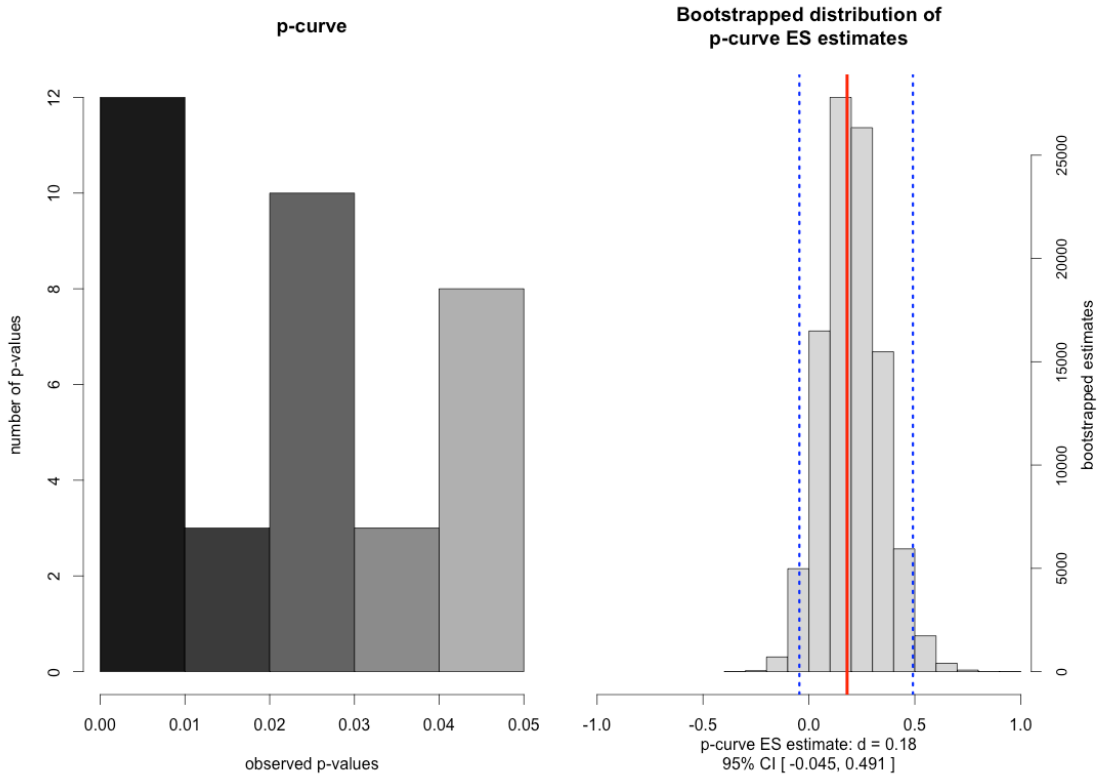

Figure 10. *p*-Curve results from HWSC dataset. Histogram of observed *p* values (the *p*-curve; left), and histogram of bootstrapped estimates of effect size simulating sampling distribution; ES estimate and 95% CI shown with solid red and dotted blue bars respectively (right).

Post-HWSC Dataset. The studies added after the cutoff date from Hagger et al. (2010) were first analyzed separately. The *p*-curve estimate of effect size for the post-HWSC dataset was $d = -0.52$, with the bootstrapped 95% confidence interval [-1.20, 0.06]. Random effects meta-analysis estimated effect size as $d = 0.69$, 95% CI [0.63,

0.76]. The *p*-curve plot for the post-HWSC dataset shows a pronounced left skew with many *p*-values between .04 and .05, strongly suggesting that studies were *p*-hacked. Typical *p*-hacking methods include (for example) collecting data for multiple dependent measures but only reporting those that 'work', or adding subjects after checking to see if results are significant; *p*-hacking is discussed in greater detail in Chapter 5.

Figure 11

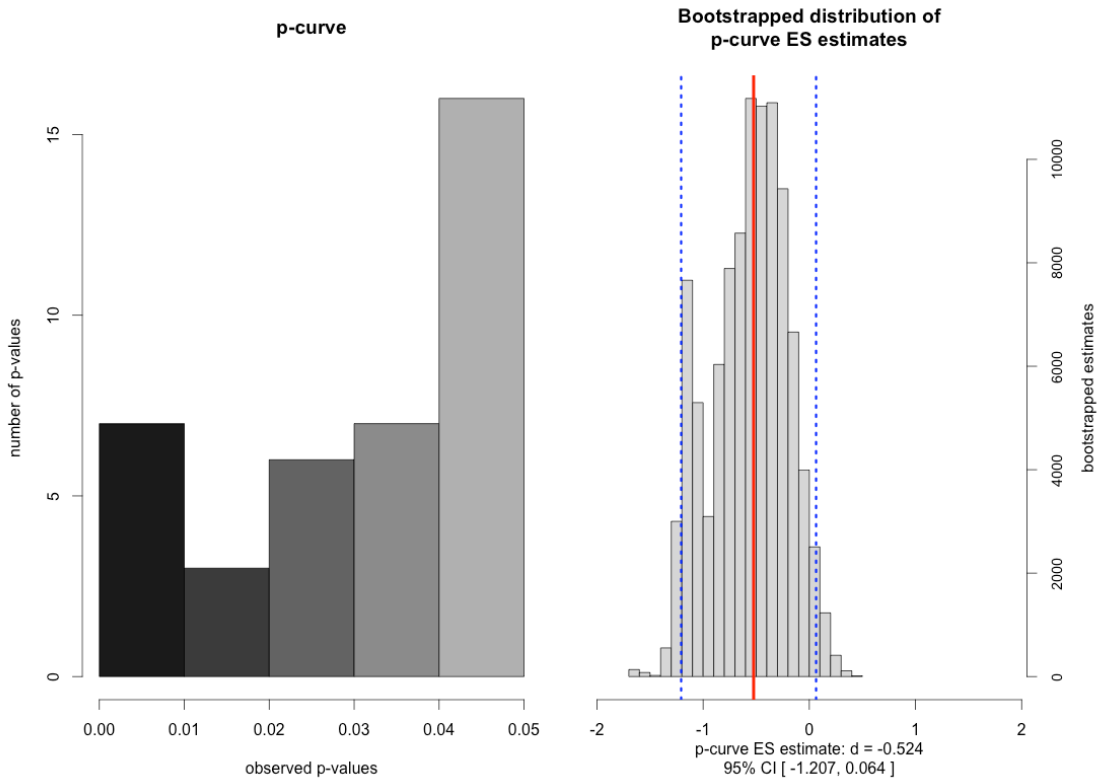

Figure 11. *p*-Curve results from the post-HWSC dataset. Histogram of observed *p* values (the *p*-curve; left). On the right is a histogram of bootstrapped estimates of effect size that simulates the sampling distribution; *p*-curve ES estimate and 95% CI shown with solid red and dotted blue bars respectively.

Full Dataset. The *p*-curve estimate of effect size for the full dataset was *d* = -0.12, with a bootstrapped 95% CI of [-0.55, 0.19]. The observed *p*-curve for the full dataset is slightly skewed to the left, consistent with the effects of *p*-hacking (Simmons, Nelson, & Simonsohn, 2011). For comparison, a random-effects meta-analysis conducted on the full dataset estimated effect size as *d* = 0.75, 95% CI [0.70, 0.80].

Figure 12

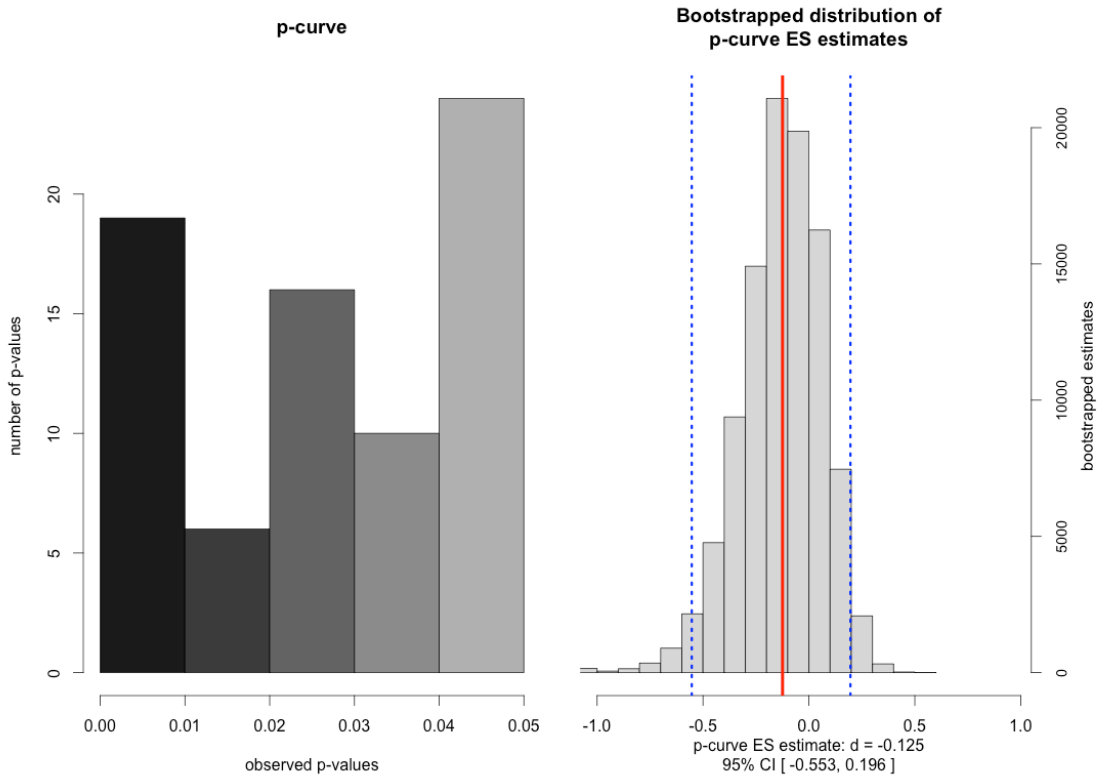

Figure 12. *p*-Curve results from full dataset. Histogram of observed *p* values (the *p*-curve; left), and histogram of bootstrapped estimates of effect size simulating sampling distribution; ES estimate and 95% CI shown with solid red and dotted blue bars respectively (right).

Chapter Five: *p*-Curve Monte Carlo Simulations

In this section, a series of Monte Carlo simulations are described that model the effects of questionable research practices (QRPs) or *p*-hacking (Simonsohn et al., 2011) on *p*-curve's performance. In these simulations, data were randomly generated on a per-subject level from known distributions to produce simulated studies, which are then subjected to the same analysis that would be used with human subjects data. These sets of simulated studies can then be subjected to meta-analysis using *p*-curve or traditional weighted mean effect size methods, and because the underlying studies were artificially generated from a known effect size, the accuracy of meta-analytic ES estimates can be measured. By simulating the behavior of a hypothetical researcher at every stage of the data collection, analysis, and publication process, we can determine the ultimate effect of various questionable practices on meta-analytic results.

Earlier simulations of this kind conducted by Simonsohn et al. (2014) suggested that *p*-curve effect size estimates were not substantially influenced by the presence of QRPs, exhibiting only a slight negative bias of approximately $d = 0.1$. However, Simonsohn et al. (2014) simulated the effects of three QRPs separately, and used especially conservative parameters for optional stopping (data peeking). It is plausible that in reality, researchers employing QRPs may use multiple methods to maximize the probability of obtaining significant (publishable) results, making it necessary to perform more naturalistic simulations.

Current Simulations

Simulating every possible combination of QRPs would result in an unmanageably large number of conditions. For the current simulations, a set of 6 hypothetical researchers (A through F) were simulated, each using different combinations of QRPs. The combinations of QRPs used by each researcher were chosen intuitively by the author; no formal procedures were employed. Four questionable research practices were simulated: optional stopping (data peeking), multiple dependent variables, elimination of 'bad' subjects, and exclusion of outliers. Details of these practices and which (simulated) researchers used which methods are shown in Table 3, and complete R code (R Core Team, 2014) used for the simulations can be found in the Online Supplement (http://dx.doi.org/10.13020/D62S33).

Table 3

Simulated Questionable Research Practice Methods

Method	Description	Used By Researchers
Optional Stopping	If *t*-test is not significant, add *n*=1 or *n*=5 new subjects to each group, then test again. Repeat until maximum sample size is reached.	B, E (*n*=5), C, F (*n*=1)
Multiple DVs	Collect data for two separate, uncorrelated dependent measures; if no significant group differences are obtained for one DV,	D, E, F

	try a second DV, and only report if significant.	
Outlier Exclusion	If *t*-test is not significant, exclude all observations >2*SD* from the group mean and test again.	D, E, F
'Bad' Subject Exclusion	If *t*-test is not significant, exclude one subject from each group whose data are least consistent with hypothesis, then run the *t*-test again.	D, E, F

All simulated studies used a simple two-cell between-subjects design and the between-subjects *t*-tests assumed equal variance. For each study that did not simulate data peeking, sample sizes were randomly chosen from a uniform distribution ranging from 10 to 40 subjects per group (group sizes were equal). For studies that did simulate data peeking, per-group starting sample sizes were drawn from a uniform distribution ranging from 10 to 20, and maximum sample sizes were drawn from a uniform distribution ranging from 30 to 40.

Studies were simulated using five different underlying true effect sizes ranging from $d = 0.0$ to $d = 0.8$. Five thousand significant studies were generated for each unique combination of researcher and underlying effect, producing a grand total of 150,000 simulated studies. Random effects meta-analyses on each set of 5000 studies were conducted using the R metafor package (Viechtbauer, 2010), in addition to *p*-curve estimates of effect size. Confidence intervals were *not* calculated in these simulations due

to the unreasonably large number of studies in each condition; the goal was simply to generate *p*-curves and associated estimates of effect size relatively free of sampling error to evaluate *p*-curve's performance under a variety of conditions. The resulting *p*-curves and effect size estimates are shown in Figure 13.

Simulation Results

Results of the simulations consisted of an ES estimate from both *p*-curve and random effects traditional MA for each possible researcher/ES combination. Researcher A does not employ any QRPs and provides a good baseline test of how well *p*-curve performs without QRPs. At each simulated effect size for researcher A, *p*-curve produces an accurate ES estimate while random effects MA does not. This shows that when QRPs are not used, *p*-curve estimates effect size very accurately even when publication bias has removed all nonsignificant findings from the analysis.

Figure 13

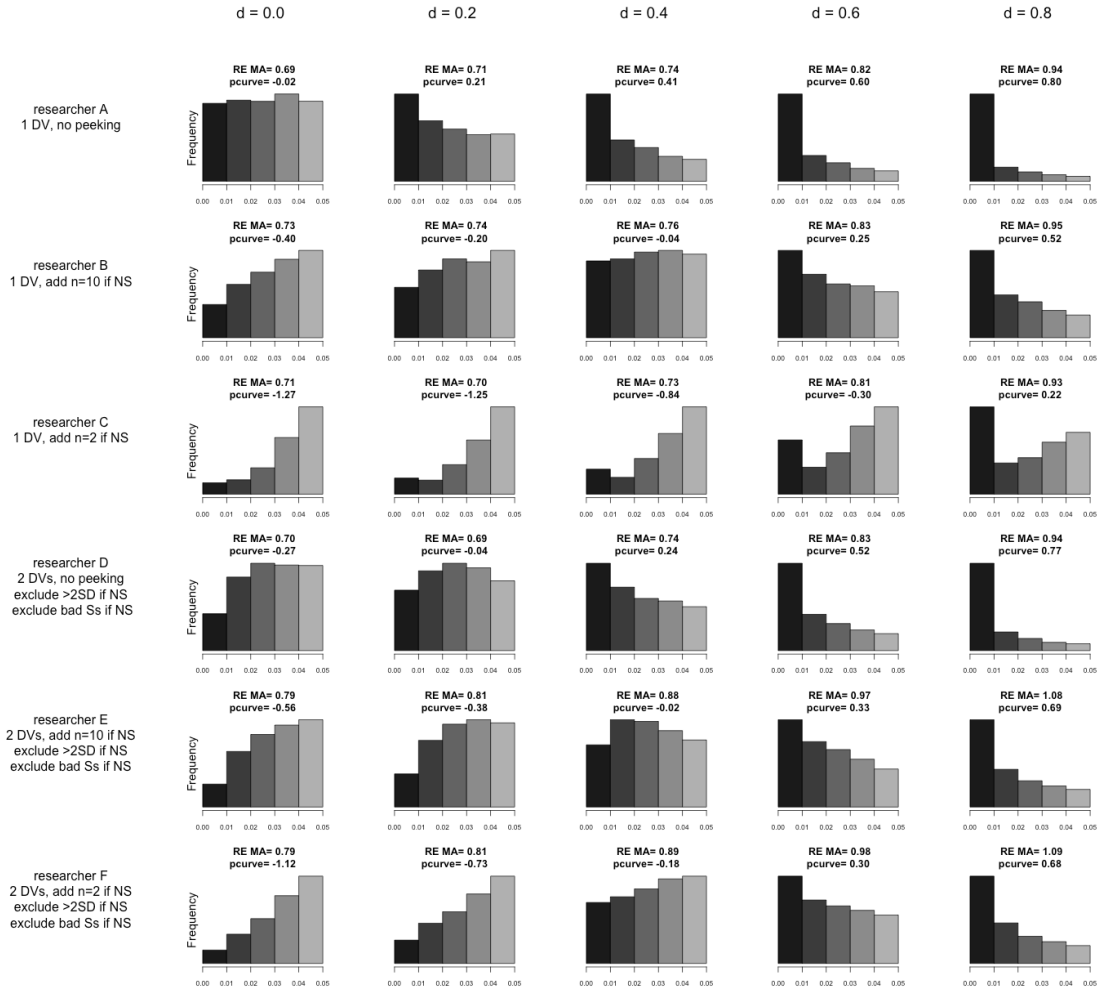

Figure 13. *p*-Curves generated by six simulated researchers (rows) at five true underlying effect sizes (columns). Each *p*-curve represents 5,000 simulated studies. RE MA = random-effects meta analysis.

When optional stopping is used in isolation as simulated by researchers B and C, *p*-curve's estimates exhibit strong negative bias, especially at small or null effect sizes.

For example, a true effect of $d = 0.4$ tested by researcher B is estimated to be null by *p*-curve. Similar results are obtained when other QRP methods (multiple DVs, outlier or subject exclusion) are employed by researchers E and F. In all cases, the use of QRPs caused *p*-curve to *under*estimate effect sizes, whereas random-effects meta-analysis *over*estimated effect sizes.

In reality, different researchers are likely to employ different QRP methods, and determining the exact combinations used in a given set of studies would likely be impossible. John et al. (2012) found the prevalence of QRPs to be near-universal in psychology, making it safe to assume that they are present in any given literature. In all cases, *p*-curve estimates for studies with an underlying true ES of $d = 0.8$ appear to be affected the least in the current simulations. Simulated studies with smaller underlying effect sizes generated *p*-curve estimates that varied markedly. For a set of studies testing small (true) effects or null effects, the present simulations show that the particular types of QRPs used (which a meta-analyst can never determine) exert some influence on *p*-curve's effect size estimate.

Overall, *p*-curve estimates of effect size are remarkably accurate when publication bias restricts available studies to those that show significant effects. *p*-Curve reliably recovers the underlying or true effect size even for sets of underpowered studies, a scenario where traditional meta-analytic methods fail. When QRPs are present, *p*-curve effect size estimates can be negatively biased; as confidence increases that a set of studies was not produced with the help of QRPs, confidence in *p*-curve effect size estimates should also increase.

Chapter Six: Discussion

The robustness of the resource depletion effect has been called into question as a result of recent replications that failed to observe depletion effects (e.g., Xu et al., 2014; Carter & McCullough, 2013b). These replications are directly at odds with meta-analytic results from Hagger et al. (2010) that suggest the resource depletion effect size is a relatively robust $d = 0.62$; it is possible that publication bias has inflated this estimate, but to what degree? The current work aims to more accurately estimate the resource depletion effect size using a behavioral experiment typical of depletion research (Behavioral Experiment) and a novel meta-analytic method, *p*-curve. Issues associated with employing *p*-curve 'in the wild', including Monte Carlo simulations, are also explored.

Empirical Results

In comparison to previously published depletion studies, the current Behavioral Experiment appears quite orthodox and robust. The silent-video depletion manipulation task used in the Behavioral Experiment (Schmeichel et al., 2003) has elicited depletion effects in many previous studies, and the Stroop task (or variations thereof) is also commonly used as a dependent measure in depletion work. A large sample size was obtained, sufficient to obtain power of 99.94% (using Hagger et al's [2010] $d = 0.62$ estimate). Despite an apparently robust design, the experiment failed to yield a significant depletion effect, and the observed effect size was $d = 0.197$; bootstrapped CIs around the observed effect size were wide enough to include zero, yet failed to include the $d = 0.62$ estimate from Hagger et al. (2010).

There are some noteworthy limitations of the behavioral study. Taken in isolation, the results suggest only that the depletion effect (for these particular tasks) is not as strong as results from Hagger et al. would suggest. Whether the true underlying effect size is effectively null or a moderately sized $d = 0.4$ is a question that results from the Behavioral Experiment alone cannot answer. In addition, only one depletion task and only one test task was used. It is possible that that particular combination of depletion and test tasks happens to be a combination that produces weaker effects compared with other combinations. However, this possibility is not supported by results from the meta-analysis indicating that similarly small effects were estimated from studies using a variety of different combinations of depleting and test tasks.

p-Curve Simulation Results

Before discussing the meta-analytic results from *p*-curve in this dissertation, it is necessary to first review the results obtained from the Monte Carlo simulations to provide some kind of context that may describe *p*-curve's strengths and limitations. When researchers are not using QRPs, *p*-curve effect size estimates are remarkably accurate at all effect sizes; publication bias has no deleterious effect whatsoever. In situations where researchers do employ QRPs, *p*-curve's estimates are negatively biased; this is most pronounced when true (underlying) effect sizes are at or below $d = 0.4$. More importantly, the degree of negative bias appears to be dependent on the types of QRPs employed. For example, researcher B tested an underlying true effect size of $d = 0.4$ and obtained an estimate of $d = -0.04$, while researcher F obtained $d = -0.18$ from the same underlying effect. Since the exact QRPs or their proportions in the available studies are

unknown in meta-analysis and the underlying effect size is unknown, it is not possible to 'work backwards' and recover an estimate that corrects for QRPs. The greater confidence one can have that QRPs are infrequent, the greater confidence one can have in *p*-curve accuracy.

Meta-Analytic Results

The overall *p*-curve estimate of effect size for depletion effects obtained here was d = -0.12 (95% CI: -0.55, 0.19), which differs considerably from the behavioral results and markedly from the estimate in Hagger et al. (2010). Examining the studies from the HWSC and post-HWSC datasets separately, large differences were apparent; *p*-curve effect size estimates from the HWSC dataset (d = 0.18) mirror almost perfectly the estimated effect size (d = 0.197) and confidence interval from the Behavioral Experiment, while the *p*-curve effect size estimate from the post-HWSC dataset (d = -0.52) was wildly negative.

Given the results from the *p*-curve simulations discussed above, the post-HWSC result could have arisen as a result of the influence of QRPs. It would be silly to suggest that the ego depletion effect size tested in studies after 2010 suddenly became d = -0.52. Rather, it seems much more reasonable to attribute results from the post-HWSC dataset to the negative influence of QRPs. This viewpoint is corroborated with a simple visual inspection of the post-HWSC *p*-curve, which exhibits a pronounced left skew (Figure 11). Distributions of *p*-values can only become skewed in this direction when QRPs are employed. This should reduce the degree to which the post-HWSC result is considered to

be informative for estimating the true effect size, notwithstanding the informative nature of possibly uncovering evidence that the use of QRPs have been on the rise since 2010.

One speculation that can be made is that the longer an effect remains popular in the literature, the greater the likelihood that researchers will tend to use QRPs to observe the effect. This is just a speculation at this point, of course, but it is one with some level of face validity. A researcher that sometimes but not always uses QRPs may feel more compelled to do so the frequently the effect appears in the literature.

p-Curve Limitations and Observations

Aside from the aforementioned problems with *p*-curve when QRPs are in use, several other issues hamper *p*-curve's utility as a meta-analytic tool. Most prominent is the requirement that limits *p*-curve to studies employing a simple two-cell design. In practice, this severely limited the number of studies that could be included in the current analyses. For example, a total of $k = 198$ studies were included in the Hagger et al. (2010) meta-analysis, but only $k = 36$ of those studies met inclusion criteria for the current study, with the majority rejected due to a lack of a 2-cell design. This is a major limitation of *p*-curve effect size estimation in comparison to more traditional methods.

In addition, an unknown number of extant two-cell studies may be the remnants of larger failed experimental designs. For example, imagine a researcher conducts a study testing the effects of a moderator on the resource depletion effect using a 2x2 design. If the key interaction does not reach significance, but the simple comparison between two cells does reach significance, the *p*-value associated with the simple comparison will be relatively large (e.g., $p = .04$). If the simple two-cell effect is not strong enough to also

cause the interaction to be significant, the associated *p*-value for the two-cell comparison will be relatively high. This is the opposite case of the *p*-curve limitation in which simple effects drawn from larger designs cannot be used; in such cases, the *p*-value distribution for the simple two-cell comparison is not uniform under the null.

Conclusions/Next Steps

Synthesizing the current empirical and meta-analytic results, I conclude that the evidence for a measurable resource depletion effect is weak at best. The empirical evidence from the Behavioral Experiment is consistent with the *p*-curve results from the HWSC dataset. Both results provide evidence against strong effect sizes such as those suggested by Hagger et al. (2010), although weak effect sizes are still a possibility. These results are consistent with results from another meta-analysis of resource depletion effects employing novel corrections for publication bias by Carter & McCullough (2014).

Figure 14

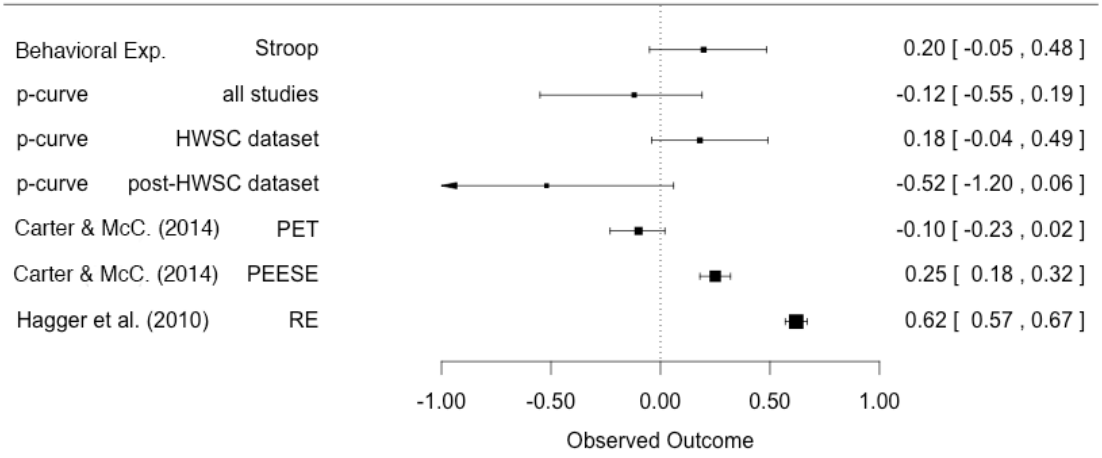

Figure 14. Forest plot comparing effect size estimates and confidence intervals from the current work with previously published estimates.

The forest plot in figure 14 conveys the uncertainty in all available estimates of the depletion effect size; it is possible that a small depletion effect exists. When smaller-sample empirical studies and meta-analytic methods are not accurate enough, the only remaining way to obtain greater precision is to conduct a large-scale replication with very large samples. The recently announced Registered Replication Reports, a specific article type published in *Perspectives on Psychological Science*, provides a vehicle for these large-scale replications.

Recent Developments

As this thesis was in preparation, results from two preregistered replications of resource depletion effects have been released. Lurquin et al. (2016) conducted a depletion study using the same manipulation as in the current Behavioral Experiment, followed by a working memory task as a dependent measure. With a total sample size of 200 subjects, Lurquin et al. (2016) found that depleted subjects exhibited working memory performance that was superior to controls, although this difference did not reach significance (Cohen's $d = 0.22$). This pattern of results, opposite that predicted by the depletion model, is similar to the results obtained by Xu et al. (2014) and Carter and McCullough (2013b).

A large-scale Registered Replication Report testing depletion effects, organized by Martin Hagger (first author of the 2010 meta-analysis), has recently finished data collection and initial results have been released. A total of 2,141 subjects in 23 labs around the world participated in the effort, which used common tasks and a typical between-subjects experimental design seen in most depletion studies discussed in this

thesis. The massive sample size obtained by Hagger et al. (2016) has provided possibly the most reliable and accurate estimate of the depletion effect size available thus far. Results from Hagger et al. (2016) did not show any evidence for a significant or even marginal depletion effect; depleted subjects performed almost identically to controls on the dependent measure, Cohen's $d = 0.04$, 95% CI: (-0.07, 0.15). This result confirms what the current Behavioral Experiment, *p*-curve analysis, and other replications and meta-analyses have suggested as to the magnitude of the resource depletion effect: *that it is effectively null*, and that published depletion studies are simply a collection of type I errors. Those who are currently conducting studies that attempt to observe a depletion effect should abandon their efforts and spend their time and effort more profitably in other areas.

References

Note: references preceded by an asterisk (*) contain studies included in the meta-analysis.

* Ackerman, J. M., Goldstein, N. J., Shapiro, J. R., & Bargh, J. A. (2009). You Wear Me Out The Vicarious Depletion of Self-Control. *Psychological Science*, 20(3), 326-332.

* 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., & Heatherton, T. F. (1996). Self-regulation failure: An overview. *Psychological Inquiry*, 7(1), 1-15.

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, 16(6), 351-355.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics,* 50(4), 1088-1101.

* Blackhart, G. C., Nelson, B. C., Winter, A., & Rockney, A. (2011). Self-control in relation to feelings of belonging and acceptance. *Self and Identity*, 10(2), 152-165.

* Bruyneel, S. D., & Dewitte, S. (2012). Engaging in self-regulation results in low-level construals. *European Journal of Social Psychology*, 42(6), 763-769.

* Bruyneel, S. D., Dewitte, S., Franses, P. H., & Dekimpe, M. G. (2009). I felt low and my purse feels light: Depleting mood regulation attempts affect risk decision making. *Journal of Behavioral Decision Making*, 22(2), 153-170.

* Bruyneel, S., Dewitte, S., Vohs, K. D., & Warlop, L. (2006). Repeated choosing increases susceptibility to affective product features. *International Journal of Research in Marketing*, 23(2), 215-225.

* Burkley, E. (2008). The role of self-control in resistance to persuasion. *Personality and Social Psychology Bulletin*, 34(3), 419–431.

Carter, E. C. (2013). *A Series of Meta-Analytic Tests of the Depletion Effect* (Doctoral dissertation). Retrieved from http://scholarlyrepository.miami.edu/oa_dissertations/1032

Carter, E. C., & McCullough, M. E. (2013a). Is ego depletion too incredible? Evidence for the overestimation of the depletion effect. *Behavioral and Brain Sciences*, 36(6), 683-684.

Carter, E. C., & McCullough, M. E. (2013b). After a pair of self-control-intensive tasks, sucrose swishing improves subsequent working memory performance. *BMC Psychology*, 1(1), 22.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated. *Frontiers in Psychology*, 5(823), 1-11.

* Ciarocco, N. J., Sommer, K. L., & Baumeister, R. F. (2001). Ostracism and ego depletion: The strains of silence. *Personality and Social Psychology Bulletin*, 27(9), 1156-1163.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

* Converse, P. D., Pathak, J., Steinhauser, E., & Homan, E. W. (2012). Repeated Self-Regulation and Asymmetric Hemispheric Activation. *Basic and Applied Social Psychology*, 34(2), 152-167.

Dear, K., & Dobson, A. (1997). Comment on Givens, G. H., Smith, DD, and Tweedie, RL (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science*, 12, 244-245.

* DeBono, A., & Muraven, M. (2013). Keeping it real: self-control depletion increases accuracy, but decreases confidence for performance. *Journal of Applied Social Psychology*, 43(4), 879-886.

* DeBono, A., Shmueli, D., & Muraven, M. (2011). Rude and inappropriate: The role of self-control in following social norms. *Personality and Social Psychology Bulletin*, 37(1), 136-146.

* Derrick, J. L. (2013). Energized by television familiar fictional worlds restore self-control. *Social Psychological and Personality Science*, 4(3), 299-307.

* DeWall, C. N., Baumeister, R. F., Stillman, T. F., & Gailliot, M. T. (2007). Violence restrained: Effects of self-regulation and its depletion on aggression. *Journal of Experimental Social Psychology*, 43(1), 62-76.

DuMouchel, W., & Harris, J. (1997). Comment on Givens, G. H., Smith, D. D., and Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science*, 12, 244-245.

Duval, S., & Tweedie, R. (2000a). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), 455-463.

Duval, S., & Tweedie, R. (2000b). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449), 89-98.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171-185.

* Egan, P. M., Hirt, E. R., & Karpen, S. C. (2012). Taking a fresh perspective: Vicarious restoration as a means of recovering self-control. *Journal of Experimental Social Psychology*, 48(2), 457-465.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629-634.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

* Fennis, B. M., & Janssen, L. (2010). Mindlessness revisited: Sequential request techniques foster compliance by draining self-control resources. *Current Psychology*, 29(3), 235-246.

* Fennis, B. M., Janssen, L., & Vohs, K. D. (2009). Acts of benevolence: A limited-resource account of compliance with charitable requests. *Journal of Consumer Research*, 35(6), 906-924.

* Finkel, E. J., Campbell, W. K., Brunell, A. B., Dalton, A. N., Scarbeck, S. J., & Chartrand, T. L. (2006). High-maintenance interaction: inefficient social coordination impairs self-regulation. *Journal of Personality and Social Psychology*, 91(3), 456-475.

* Fischer, P., Kastenmüller, A., & Asal, K. (2012). Ego depletion increases risk-taking. *The Journal of Social Psychology*, 152(5), 623-638.

* Fischer, P., Greitemeyer, T., & Frey, D. (2008). Self-regulation and selective exposure: the impact of depleted self-regulation resources on confirmatory information processing. *Journal of Personality and Social Psychology*, 94(3), 382-395.

Fisher, R. A. (1932). *Statistical Methods for Research Workers (4th Ed.)*. Oxford, England: Oliver & Boyd.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.

* Freeman, N., & Muraven, M. (2010). Self-control depletion leads to increased risk taking. *Social Psychological and Personality Science*, 1(2), 175-181.

Freud, S. (1920). *A general introduction to psychoanalysis*. New York, NY: Horace Liveright.

* Furley, P., Bertrams, A., Englert, C., & Delphia, A. (2013). Ego depletion, attentional control, and decision making in sport. *Psychology of Sport and Ex*ercise, 14(6), 900-904.

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., ... & Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy

source: willpower is more than a metaphor. *Journal of Personality and Social Psychology*, 92(2), 325-336.

* Gailliot, M. T., Gitter, S. A., Baker, M. D., & Baumeister, R. F. (2012). Breaking the rules: Low trait or state self-control increases social norm violations. *Psychology*, 3(12), 1074-1083.

* 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.

* Geeraert, N., & Yzerbyt, V. Y. (2007). How fatiguing is dispositional suppression? Disentangling the effects of procedural rebound and ego-depletion. *European Journal of Social Psychology*, 37(2), 216-230.

Gilbert, D. T., Krull, D. S., & Pelham, B. W. (1988). Of thoughts unspoken: Social inference and the self-regulation of behavior. *Journal of Personality and Social Psychology*, 55(5), 685-694.

* Martin Ginis, K. A., & Bray, S. R. (2010). Application of the limited strength model of self-regulation to understanding exercise effort, planning and adherence. *Psychology and Health*, 25(10), 1147-1160.

Givens, G. H., Smith, D. D., & Tweedie, R. L. (1997). Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, 12(4), 221-240.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3-8.

* Gotlib, T., & Converse, P. (2010). Dishonest Behavior: The Impact of Prior Self-Regulatory Exertion and Personality. *Journal of Applied Social Psychology*, 40(12), 3169-3191.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464-1480.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., ... Zwienenberg, M. (2016). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*. Retrieved from http://www.psychologicalscience.org/index.php/publications/rrr-the-ego-depletion-paradigm/.

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.

* Healey, M. K., Hasher, L., & Danilova, E. (2011). The stability of working memory: Do previous tasks influence complex span? *Journal of Experimental Psychology: General*, 140(4), 573-585.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2), 246-255.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486-504.

Inzlicht, M., & Gutsell, J. N. (2007). Running on empty neural signals for self-control failure. *Psychological Science*, 18(11), 933-937.

Inzlicht, M., & Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, 7(5), 450-463.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3(1), 109-117.

James, W. (1890). *The principles of psychology (Vol. 1)*. New York, NY: Holt.

* Janssen, L., Fennis, B. M., Pruyn, A. T. H., & Vohs, K. D. (2008). The path of least resistance: Regulatory resource depletion and the effectiveness of social influence techniques. *Journal of Business Research*, 61(10), 1041-1045.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524-532.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65(1), 51-69.

* Kim, J., Kim, J. E., & Park, J. (2012). Effects of cognitive resource availability on consumer decisions involving counterfeit products: The role of perceived justification. *Marketing Letters*, 23(3), 869-881.

Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, 45(4), 905-927.

Kurzban, R. (2010). Does the brain consume additional glucose during self-control tasks? *Evolutionary Psychology*, 8(2), 244-259.

* Li, J. B., Nie, Y. G., Zeng, M. X., Huntoon, M., & Smith, J. L. (2013). Too exhausted to remember: Ego depletion undermines subsequent event-based prospective memory. *International Journal of Psychology*, 48(6), 1303-1312.

Light, R., & Smith, P. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 41(4), 429-471.

Lurquin, J.H., Michaelson, L.E., Barker, J.E., Gustavson, D.E., von Bastian, C.C., Carruth, N.P., et al. (2016). No Evidence of the Ego-Depletion Effect across Task Characteristics and Individual Differences: A Pre-Registered Study. *PLoS ONE*, 11(2), e0147770.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, 109(2), 163-203.

McCracken, M. O. (2015). BibDesk (version 1.6.4). Retrieved from http://bibdesk.sourceforge.net

* McEwan, D., Martin Ginis, K. A., & Bray, S. R. (2013). The effects of depleted self-control strength on skill-based task performance. *Journal of Sport and Exercise Psychology*, 35(3), 239-249.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103-115.

* Milkman, K. L. (2012). Unsure what the future will bring? You may overindulge: Uncertainty increases the appeal of wants over shoulds. *Organizational Behavior and Human Decision Processes*, 119(2), 163-176.

* Molden, D. C., Hui, C. M., Scholer, A. A., Meier, B. P., Noreen, E. E., D'Agostino, P. R., & Martin, V. (2012). Motivational versus metabolic effects of carbohydrates on self-control. *Psychological Science*, 23(10), 1137-1144.

* Molet, M., Miller, H. C., Laude, J. R., Kirk, C., Manning, B., & Zentall, T. R. (2012). Decision making by humans in a behavioral task: Do humans, like pigeons, show suboptimal choice? *Learning & behavior*, 40(4), 439-447.

Mosteller, F., Bush, R. R., & Green, B. F. (1954). *Selected quantitative techniques*. Boston, MA: Addison-Wesley.

* Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as a limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology*, 74(3), 774-789.

Oaten, M., & Cheng, K. (2006). Improved self-control: The benefits of a regular program of academic study. *Basic and Applied Social Psychology*, 28(1), 1-16.

Oaten, M., & Cheng, K. (2007). Improvements in self-control from financial monitoring. *Journal of Economic Psychology*, 28(4), 487-501.

* Oaten, M., Williams, K. D., Jones, A., & Zadro, L. (2008). The effects of ostracism on self-regulation in the socially anxious. *Journal of Social and Clinical Psychology*, 27(5), 471-504.

Olkin, I. (1995). Statistical and theoretical considerations in meta-analysis. *Journal of Clinical Epidemiology*, 48(1), 133-146.

* Ostafin, B. D., Marlatt, G. A., & Greenwald, A. G. (2008). Drinking without thinking: An implicit measure of alcohol motivation predicts failure to control alcohol use. *Behaviour Research and Therapy*, 46(11), 1210-1219.

R Core Team (2014). R: A language and environment for statistical computing (version 3.1.2). R Foundation for Statistical Computing, Vienna, Austria.

Richeson, J. A., & Shelton, J. N. (2003). When prejudice does not pay effects of interracial contact on executive function. *Psychological Science*, 14(3), 287-290.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-641.

RStudio Team (2015). RStudio: Integrated Development for R (version 0.99.489). RStudio, Inc., Boston, MA.

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.

* Schmeichel, B. J., Vohs, K. D., & Baumeister, R. F. (2003). Intellectual performance and ego depletion: role of the self in logical reasoning and other information processing. *Journal of Personality and Social Psychology,* 85(1), 33-46.

* Schmeichel, B. J., Harmon-Jones, C., & Harmon-Jones, E. (2010). Exercising self-control increases approach motivation. *Journal of Personality and Social Psychology*, 99(1), 162-173.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.

Sedikides, C., Campbell, W. K., Reeder, G. D., & Elliot, A. J. (1998). The self-serving bias in relational context. *Journal of Personality and Social Psy*chology, 74(2), 378.

Shadish, W. R., & Lecy, J. D. (2015). The meta-analytic big bang. *Research Synthesis Methods*, 6(3), 246-264.

Shamosh, N. A., & Gray, J. R. (2007). The relation between fluid intelligence and self-regulatory depletion. *Cognition and Emotion*, 21(8), 1833-1843.

* Shmueli, D., & Prochaska, J. J. (2009). Resisting tempting foods and smoking behavior: implications from a self-control theory perspective. *Health Psychology*, 28(3), 300-306.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-Curve and Effect Size Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*, 9(6), 666-681.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Associa*tion, 54(285), 30-34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician,* 49(1), 108-112.

* Stucke, T. S., & Baumeister, R. F. (2006). Ego depletion and aggressive behavior: Is the inhibition of aggression a limited resource? *European Journal of Social Psychology*, 36(1), 1-13.

Tyler, J. M., & Burns, K. C. (2008). After depletion: The replenishment of the self's regulatory resources. *Self and Identity*, 7(3), 305-321.

Underwood, B. J. (1957). Interference and forgetting. *Psychological review*, 64(1), 49-60.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48.

Villar, J., Piaggio, G., Carroli, G., & Donner, A. (1997). Factors affecting the comparability of meta-analyses and largest trials results in perinatology. *Journal of Clinical Epidemiology*, 50(9), 997-1002.

* Vohs, K. D., Baumeister, R. F., & Ciarocco, N. J. (2005). Self-regulation and self-presentation: regulatory resource depletion impairs impression management and effortful self-presentation depletes regulatory resources. *Journal of Personality and Social Psychology*, 88(4), 632-657.

* 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.

* Vohs, K. D., & Faber, R. J. (2007). Spent resources: Self-regulatory resource availability affects impulse buying. *Journal of Consumer Research*, 33(4), 537-547.

* Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *Psychological Science*, 11(3), 249-254.

* Vohs, K. D., & Schmeichel, B. J. (2003). Self-regulation and extended now: Controlling the self alters the subjective experience of time. *Journal of Personality and Social Psychology*, 85(2), 217-230.

Von Hippel, W., & Henry, J. D. (2011). Aging and self-regulation. In Vohs, K. D., Baumeister, R. F. (Ed.), Handbook of Self-Regulation (pp. 321). New York, NY: Guilford Press.

* Wan, E. W., & Agrawal, N. (2011). Carryover effects of self-control on decision making: A construal-level perspective. *Journal of Consumer Research*, 38(1), 199-214.

* Wan, E. W., Rucker, D. D., Tormala, Z. L., & Clarkson, J. J. (2010). The effect of regulatory depletion on attitude certainty. *Journal of Marketing Research*, 47(3), 531-541.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, 53(1), 5-13.

Wickham, H., & Francois, R. (2015). dplyr: A Grammar of Data Manipulation (version 0.4.1.). Retrieved from http://CRAN.R-project.org/package=dplyr

Xu, X., Demos, K. E., Leahey, T. M., Hart, C. N., Trautvetter, J., Coward, P., ... & Wing, R. R. (2014). Failure to replicate depletion of self-control. *PloS One*, 9(10), 1-5.

* Xu, H., Bègue, L., & Bushman, B. J. (2012). Too fatigued to care: Ego depletion, guilt, and prosocial behavior. *Journal of Experimental Social Psychology*, 48(5), 1183-1186.

Appendix A

Table 1

Wordlists used in Experiment 1 video task, sampled from Francis and Kucera (1982)

List 1	List 2	List 3	List 4
GOOSE	SIMPLY	TOURIST	TREE
MANNER	ELEMENT	GARDEN	ROUTE
FANG	FLAT	SECTION	PHOTO
SURVIVE	DENTAL	CURIOUS	DRAMA
SINK	RENDER	CLOCK	ARMAMENT
ADMIT	WITCH	VICE	HALF
UNSCREW	BEAR	SCOOP	SORORITY
SUNRISE	MARBLE	RAMBLE	COPY
CAPSULE	WOOD	CRISIS	SWEEP
GAME	SWITCH	PINK	MISS
PULPIT	DEBATE	FALL	ABANDON
ALLEGE	ANNALS	POLICY	TEND
SPIKE	ACCUSE	DETERGENT	AFFECT
BAND	BATTLE	PACK	POTION
DEMAND	BUSY	ASSIST	PURPOSE
SLUMP	LEAD	ALTER	LAMB
VISITOR	STICK	RANGE	UNION
TARGET	FAIRY	SPARK	MELT
LATER	APPLICATION	DEATH	READ
MEDAL	WINK	REFER	EMBED
TANGENT	PASS	DELEGATE	MERE
INSIST	SLOW	MENTAL	SILLY
FETCH	REDHEAD	MALE	SUPPOSE
HOUSE	TEAM	MAGNET	WILLOW

Appendix B

Figure 1: Subject instruction sheet

CCI-VID-70 Experiment ***FORM A***

In this experiment, you are going to be completing a variety of different tasks that combine the interests of several different researchers and research areas. Below are the instructions for each task; the experimenter will let you know when to perform each task. Throughout the experiment, please keep your chin on the chinrest so that your eyes remain a constant distance from the computer display.

Video Task:

In the video task, you will watch a silent video of a person being interviewed. After the video, you'll be asked to fill out some questionnaires on the personality of the person being interviewed. You will also see some words on the bottom of the screen – it is important for your condition of the experiment that you keep your eyes focused on the participant's face during the video and do not under any circumstances look down at the words that are appearing at the bottom of the screen. If you do accidentally look at one of the words, look away as quickly as possible. During this task, please remember to keep your chin in the chinrest to maintain a consistent distance from the monitor.

Color/Word Task:

1) First, a small white cross will appear on the screen for a half second. Please look at this cross as long as it appears on the screen.

2) Then, a word will appear on the screen. The letters of the word will be printed in one of two colors, green or purple.

3) Then, you should press one of two buttons. If the letters are colored purple, press the LEFT button using the index finger of your right hand. If the letters are colored green, press the RIGHT button using the middle finger of your right hand. Please respond as quickly and as accurately as you can. Please rest your fingers on those response buttons throughout the task in order to help you respond quickly. It is very important that your button press reflects the color in which the letters appear, not the word that appears. For example, if the word "GREEN" appears to you with purple letters, you should press the purple response button, not the green response button. After you press a button, the next trial will begin automatically one second later.

Before the experiment begins, you will briefly practice the video and color/word tasks in order to familiarize yourself with them.

If you have any questions, please ask the researcher.

Appendix C

In order to find depletion studies published after the cutoff date of April 1st, 2009 in the Hagger et al. (2010) meta-analysis, keyword and reference searches were conducted on three separate abstract search engines. These searches produced three separate sets of references (with abstracts included); the sets were merged and duplicate entries were removed, producing a total of 6,324 articles. Queries performed on two of the three electronic search engines (Ovid and Web of Science) were configured to only return articles citing any of three early publications on depletion effects (Baumeister et al., 1998; Muraven et al., 1998; and Muraven & Baumeister, 2000). This restriction served to reduce the number of erroneous results while preserving publications of potential interest, all of which should contain references to early or 'classic' depletion studies. A similar strategy to reduce erroneous results was used by Carter (2013). Details of the searches performed on each search engine are listed below.

Queries submitted to Ovid searched the MEDLINE, Embase, and PsycINFO databases. Three separate searches were conducted on publications citing the three depletion studies mentioned above using the following search string: ("Self regulat*" or "resource deplet*" or "depleted resource*" or "ego depletion" or "limited resource*" or "regulatory resource*"). Search results were limited to articles published between 2009 and 2013 in the English language. A total of 1,606 unique publications were returned from the Ovid queries.

The single query submitted to EBSCOHost searched the ERIC database only; search results were *not* limited to articles citing any of the three 'classic' depletion studies

mentioned above. The following search string was used: ("Self regulat*" or "resource deplet*" or "depleted resource*" or "ego depletion" or "limited resource*" or "regulatory resource*"). Search results were limited to peer-reviewed journal articles published between April 2009 and December 2013. A total of 1,142 unique publications were returned from the EBSCOHost query.

Queries submitted to Web of Science were 'refined' using the following criteria:

> WEB OF SCIENCE CATEGORIES: (PSYCHOLOGY MULTIDISCIPLINARY OR EDUCATION EDUCATIONAL RESEARCH OR PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR PSYCHOLOGY DEVELOPMENTAL OR SOCIAL SCIENCES INTERDISCIPLINARY OR PSYCHOLOGY SOCIAL OR PSYCHOLOGY CLINICAL OR PSYCHOLOGY EDUCATIONAL OR PSYCHOLOGY OR PSYCHOLOGY EXPERIMENTAL OR NEUROSCIENCES OR PSYCHIATRY OR PSYCHOLOGY APPLIED OR MANAGEMENT OR BUSINESS OR MULTIDISCIPLINARY SCIENCES OR BEHAVIORAL SCIENCES OR CLINICAL NEUROLOGY) AND DOCUMENT TYPES: (ARTICLE)

Three separate searches were conducted on publications citing the three 'classic' depletion studies mentioned above with no search string, in addition to a single general search (citations ignored) using the following search string: ("Self regulat*" or "resource deplet*" or "depleted resource*" or "ego depletion" or "limited resource*" or "regulatory resource*"). Search results were limited to articles published between 2009 and 2013. A total of 5,051 unique publications were returned from the Web of Science queries.

Appendix D

Abbreviated p-*curve Disclosure Table. Full version accessible in online supplement located at (http://dx.doi.org/10.13020/D62S33).*

Year	Authors	Experiment	df	t-value
1998	Baumeister, Bratslavsky, et al.	3	28	2.12
2006	Bruyneel, Dewitte, et al.	2	42	2.11
2008	Burkley	3	76	2.01
2008	Burkley	1	71	1.96
2008	Burkley	2	20	2.12
2001	Ciarocco, Sommer, et al.	2	22	2.30
2007	DeWall, Baumeister, et al.	4	95	2.54
2007	DeWall, Baumeister, et al.	1	38	2.07
2009	Fennis, Janssen, et al.	2	58	2.27
2009	Fennis, Janssen, et al.	1	37	2.11
2006	Finkel, Campbell, et al.	3	44	2.68
2006	Finkel, Campbell, et al.	5	27	2.85
2006	Finkel, Campbell, et al.	1	24	2.31
2008	Fischer, Greitemeyer, et al.	1	47	2.17
2008	Fischer, Greitemeyer, et al.	4	46	2.97
2006	Gailliot, Schmeichel, et al.	3	65	1.99
2006	Gailliot, Schmeichel, et al.	2	17	2.15
2006	Geeraert & Yzerbyt	2	30	3.16
2008	Janssen, Fennis, et al.	1	58	2.25
1998	Muraven, Tice, et al.	3	47	2.07
2008	Oaten, Williams, et al.	2	72	2.68
2008	Ostafin, Marlatt, et al.	1	83	1.99
2007	Schmeichel	4	63	3.18
2003	Schmeichel, Vohs, et al.	2	35	2.27
2003	Schmeichel, Vohs, et al.	1	22	3.87
2006	Stucke & Baumeister	2	60	2.31
2006	Stucke & Baumeister	1	58	3.11
2006	Stucke & Baumeister	3	43	3.94
2007	Vohs & Faber	1	33	2.82
2008	Vohs, Baumeister, et al.	4.2	38	2.24
2008	Vohs, Baumeister, et al.	4.1	38	2.67
2008	Vohs, Baumeister, et al.	1.1	28	3.68
2008	Vohs, Baumeister, et al.	1.2	28	2.77

Year	Authors	Experiment	df	t-value
2008	Vohs, Baumeister, et al.	2	23	2.44
2008	Vohs, Baumeister, et al.	3	22	2.43
2005	Vohs, Baumeister, et al.	5	33	2.34
2000	Vohs & Heatherton	3	34	2.27
2000	Vohs & Heatherton	2	26	2.04
2003	Vohs & Schmeichel	3	46	4.05
2003	Vohs & Schmeichel	4	44	2.25
2009	Ackerman, Goldstein, et al.	1	56	2.22
2011	Blackhart, Nelson, et al.	2	53	2.22
2012	Bruyneel & Dewitte	1.1	38	2.62
2012	Bruyneel & Dewitte	1.2	61	2.72
2012	Bruyneel & Dewitte	2	61	2.06
2012	Bruyneel & Dewitte	3	104	2.06
2009	Bruyneel, Dewitte, et al.	3	24	2.38
2012	Converse, Pathak, et al.	2	16	2.50
2012	Converse, Pathak, et al.	1	39	2.03
2012	Converse, Pathak, et al.	4	43	2.10
2012	Converse, Pathak, et al.	3	92	2.03
2013	DeBono & Muraven	1	56	2.16
2013	DeBono & Muraven	2	65	4.24
2011	DeBono, Shmueli, et al.	2	34	3.52
2012	Derrick	0	44	2.68
2012	Egan, Hirt, et al.	3	25	3.13
2010	Fennis & Janssen	1	54	2.21
2009	Fennis, Janssen, et al.	1	37	2.11
2009	Fennis, Janssen, et al.	2	58	2.27
2012	Fischer, Kastenmueller, et al.	2	28	2.14
2012	Fischer, Kastenmueller, et al.	4	35	2.11
2010	Freeman & Muraven	2	44	2.04
2010	Freeman & Muraven	1	68	2.25
2013	Furley, Bertrams, et al.	1	38	2.32
2012	Gailliot, Gitter, et al.	1	43	2.03
2010	Ginis, Bray, et al.	1	59	2.07
2010	Gotlib & Converse	1	53	1.03
2011	Healey, Hasher, et al.	3	35	2.17
2011	Healey, Hasher, et al.	1	36	4.06
2011	Healey, Hasher, et al.	6	46	1.17
2011	Healey, Hasher, et al.	4	47	0.43
2011	Healey, Hasher, et al.	2	48	0.40
2011	Healey, Hasher, et al.	5	48	1.26

Year	Authors	Experiment	df	t-value
2012	Kim, Kim, et al.	1	56	2.03
2013	Li, Nie, et al.	1	38	3.28
2013	McEwan, Ginis, et al.	1	60	2.13
2012	Milkman	1	149	1.94
2012	Molden, Hui, et al.	1	83	2.09
2012	Molet, Miller, et al.	2	28	2.05
2010	Schmeichel, Harmon-Jones, et al.	1	39	2.11
2009	Shmueli & Prochaska	1	95	2.24
2011	Wan & Agrawal	1	41	2.10
2010	Wan, Rucker, et al.	1	52	3.08
2010	Wan, Rucker, et al.	2	53	2.04
2012	Xu, Begue, et al.	1	45	2.42