University of Miami
Scholarly Repository

Open Access Dissertations | Electronic Theses and Dissertations

2013-06-18

A Series of Meta-analytic Tests of the Depletion Effect

Evan C. Carter
University of Miami, evan.c.carter@gmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
Carter, Evan C., "A Series of Meta-analytic Tests of the Depletion Effect" (2013). *Open Access Dissertations*. 1032.
https://scholarlyrepository.miami.edu/oa_dissertations/1032

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.

UNIVERSITY OF MIAMI

A SERIES OF META-ANALYTIC TESTS OF THE DEPLETION EFFECT

By

Evan C. Carter

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

June 2013

©2013
Evan C. Carter
All Rights Reserved

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

A SERIES OF META-ANALYTIC TESTS OF THE DEPLETION EFFECT

Evan C. Carter

Approved:

_____ Michael E. McCullough, Ph.D. Professor of Psychology	_____ M. Brian Blake, Ph.D. Dean of the Graduate School
_____ Charles S. Carver, Ph.D. Professor of Psychology	_____ Amishi P. Jha, Ph.D. Associate Professor of Psychology
_____ Kiara R. Timpano, Ph.D. Assistant Professor of Psychology	_____ Soyeon Ahn, Ph.D. Assistant Professor, Educational and Psychological Studies

CARTER, EVAN, C. (Ph.D., Psychology)

A Series of Meta-analytic Tests of the Depletion Effect.

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Michael E. McCullough.
No. of pages in text. (113)

Few models of self-control have generated as much scientific interest as the limited strength model. The basic pattern of results predicted by this model is that acts of self-control that follow previous acts of self-control will be less likely to succeed (i.e., the so-called depletion effect). Based on results from a recent meta-analysis, researchers have concluded that the depletion effect is robust across experimental contexts and consistently medium in magnitude. Here, I detail three reasons to think that these estimates are inflated. To correct these estimates, I updated the earlier meta-analytic dataset and applied a set of statistical analyses to assess and correct for small-study effects, such as publication bias. Generally, strong signals of publication bias were found, as well as other possible small-study effects. When these influences were corrected for, there was little evidence of an effect that was distinguishable from zero. I discuss my results in terms of support for the depletion effect as proposed in the limited strength model, and I conclude that, until greater certainty about the existence of the depletion effect can be established, circumspection about the existence of this phenomenon is warranted.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Self-regulation, or the ability to alter one's state in response to perceived discrepancies between desired states and actual states, is thought to underlie a wide range of important personal and social phenomena. Better self-regulation, broadly defined and measured through a variety of techniques, has been related to greater scholastic achievement (Feldman, Martinez-pons, & Shaham, 1995; Wolfe & Johnson, 1995; Tangney, Baumeister, & Boone, 2004), higher-quality interpersonal relationships (Davis & Oathout, 1987; Mischel, Shoda, & Peake, 1988; Shoda, Mischel, & Peake, 1990; Tangney, Baumeister, & Boone, 2004), less violence and aggression (Latham & Perlow, 1996; Nigg, Quamma, Greenberg, & Kusche, 1999), a greater tendency to save money (Romal & Kaplan, 1995), and a lower chance of committing suicide (Mann, Waternaux, Gretcher, & Malone, 1999). Additionally, deficient self-regulation is thought to be an important aspect of criminality (Gottfredson & Hirschi, 1990; Burton, Cullen, Evans, Alarid, & Dunaway, 1998; McGuire & Broomfield, 1994) and the cognitive processes that underlie a variety of mental disorders, such as attention-deficit/hyperactivity disorder, bipolar disorder, substance abuse and dependence, pathological gambling, and personality disorders (Moeller, Barratt, Dougherty, Schmitz, & Swann, 2001; Madden & Bickel, 2010; American Psychiatric Association, 1994). Self-regulation is also a topic that stands out for the disciplinary breadth of the scholars who have taken an interest in it. The topic has been addressed by many sub-fields of psychology, including social-personality, clinical, developmental, health, cognitive, and evolutionary psychology, as well as fields outside of Psychology, such as business and management, biology, education, neuroscience, and computer science.

Given the apparent practical and scientific importance of self-regulation, it is unsurprising that a number of theories have been proposed to help explain how self-regulation operates and how it obtains its associations with such a range of outcomes. Of the various explanatory models of self-regulation, one of the most influential and highly publicized is the limited strength model (Baumeister & Muraven, 2000; Baumeister, Vohs, & Tice, 2007; Bauer & Baumeister, 2011). Two critical points specified by this model are (a) that volitional actions, such as acts of self-regulation, all require the expenditure of a limited resource, and (b) that to the extent that this resource is unavailable, such actions will not be successfully produced.

In this dissertation, I conducted a series of meta-analyses on the empirical findings available to date that are thought to empirically evaluate the limited strength model. The following sections describe definitions of the important terms in the limited strength model, as well as the current state of the evidence for the limited strength model. The vast majority of laboratory findings that are thought to support the limited strength model use a specific laboratory paradigm in which participants are given a sequence of at least two tasks. To date, hundreds of effect sizes derived from this experimental method have been published, and in 2010, Hagger, Wood, Stiff, and Chatzisarantis presented a meta-analysis of 83 published experiments (comprising 198 effect sizes). I will discuss Hagger et al.'s meta-analytic methods and results at length, as they form the foundation for this project, as well as highlight the difficulties in synthesizing the experimental tests of this influential theory.

Definitions

Because this topic is so widely studied, foundational terms such as self-regulation and self-control take on different meanings for different scholars. Here, I lay out explicit definitions of the important constructs in the limited strength model.

Self-regulation. The primary theorists of the limited strength model generally use some version of the following definition for self-regulation:

> Self-control or self-regulation (terms that we use interchangeably) is defined as the capacity to override natural and automatic tendencies, desires, or behaviors; to pursue long-term goals, even at the expense of short-term attractions; and to follow socially prescribed norms and rules. In other words, self-regulation is the capacity to alter the self's responses to achieve a desired state or outcome that otherwise would not arise naturally. Thus, the goal of self-control is to interrupt the self's tendency to operate on automatic pilot and to steer behavior consciously in a desired direction (p. 65; Bauer & Baumeister, 2011).

As is apparent from this quotation, the term self-regulation (or self-control) covers a lot of conceptual ground. Some scholars apply the term self-regulation to purposeful processes in general, whether effortless and automatic or conscious and deliberate, and reserve the term self-control for specific instances in which enacting one action requires *overriding* some action tendency (Carver & Scheier, 2011). This distinction is not made in the limited strength model: An act of self-regulation is not necessarily effortful, nor does it necessarily involve one process overriding another. Instead, to qualify as an act of self-regulation, an action need only involve the active alteration of automatic responses by the self.

The limited strength model, self-regulatory strength, and ego depletion. Like other theories of self-regulation, the limited strength model makes use of the concept of a

negative feedback loop (Carver & Scheier, 1982; Powers, 1973; Wiener, 1948). In principle, feedback loops include three parts: (1) *a reference value*, some set value of what the state of the system ought to be, (2) *a comparator*, a process of some form which compares the current state of the system to the reference value, and (3) *an output function*, which, when there is a discrepancy between the current state and the reference value, does something to lessen that discrepancy. The feedback loop is called "negative" because it reduces discrepancies between the current state and the reference value. The negative feedback loop that is described in the limited strength model takes its basic form from Carver and Scheier (1982, 1998) and includes: (1) *standards or goals*, abstract conceptualizations of desired states, or how the self ought to be; (2) *monitoring*, a process of some form, such as self-focused attention, that gives the self information about the current status of the self in relation to its standards and goals; and (3) an *output function*, or the means for the self to operate on itself to lessen the discrepancy between the current state and the desired state (i.e., the standard).

The limited strength model includes these three elements of the feedback loop, but differs from other theories of self-regulation in that it adds one additional feature: Successful self-regulation is thought to depend on a limited supply of *self-regulatory strength*. Within the feedback loop, once a discrepancy has been detected, a person must take action to correct it, and the level of self-regulatory strength available determines how successful the person will be in bringing his or her self back into line with the standard. For example, if a person had the goal of staying focused on a task (a standard), but found him or herself being distracted (a discrepancy), then he or she would need to apply self-regulatory strength to re-focus his or her attention.

The reduction of discrepancies, which requires the application of strength by the self, is thought to be analogous to the physical application of strength inasmuch as physical strength is limited and appears to become "depleted." The principal hypothesis of the limited strength model is that the active alteration of automatic responses depends on the limited resource of self-regulatory strength, so that any action requiring the resource depletes the current store of the resource and leaves less of it for subsequent use (Baumeister & Muraven, 2000; Baumeister, Vohs, & Tice, 2007; Bauer & Baumeister, 2011). I will refer to this as the *resource depletion hypothesis*. A lack of the necessary resource is characterized by a state of so-called *ego depletion*: "The core idea behind ego depletion is that the self's acts of volition draw on some limited resource, akin to strength or energy and that, therefore, one act of volition will have a detrimental impact on subsequent volition" (Baumeister, Bratslavsky, Muraven, & Tice, 1998). And importantly, "…the depletion of self-regulatory resources is contingent upon the operations of the active but not the passive self" (Bauer & Baumeister, 2011), so that habitual or automatic acts do not drain the limited resource. In its broadest form, the limited strength model is not simply a theory of self-regulation, but really a theory about the active, intentional self and the consequences of volition.

The sequential task paradigm. To date, nearly all of the evidence bearing on the resource depletion hypothesis is derived from an experimental sequential task paradigm. This design always includes at least two tasks. The first task is used as the independent variable and it varies by the amount of self-regulation strength required: Participants in the experimental condition are given a version of the task that is supposed to require more self-regulatory strength to perform than the version of the task given to participants

in the control condition. Following the first task, though often not immediately, a second task that also requires self-regulatory strength is given to all participants. Performance on this second task is used as the dependent variable. Researchers claim support for the resource depletion hypothesis when performance on the second task is lower for participants in the experimental condition (i.e., those participants who initially exerted greater use of the active self) compared to the performance of those in the control condition, because this condition-specific decrease in performance implies that participants' self-regulatory resources were depleted by the first task. I will refer to this pattern of results as the *depletion effect*.

As Baumeister, Brastlavsky, Muraven, and Tice noted, the depletion effect should be present as long as the outcome task and the version of the task performed by participants in the experimental condition both require use of the active self, regardless of how different those tasks appear to be:

> Our research strategy was to look at effects that would carry over across wide gaps of seeming irrelevance. If resisting the temptation to eat chocolate can leave a person prone to give up faster on a difficult, frustrating puzzle, that would suggest that those two very different acts of self-control draw on the same limited resource. And if making a choice about whether to make a speech contrary to one's opinions were to have the same effect, it would suggest that that very same resource is also the one used in general for deliberate, responsible decision making (p. 1252).

Furthermore, performing an initial task that requires use of the active self should induce no change in performance on a subsequent task that requires little use of the active self, such as providing the definitions of words (Gailliot, Schmeichel, & Baumeister, 2006). Therefore, the key aspect of the design of the sequential task paradigm is that the features of the tasks included in the paradigm can be systematically varied so that one may test

whether performance decrements exist *only* for participants in the experimental condition and *only* on outcome tasks that require use of the active self.

There are two important consequences of the reliance on this paradigm for the limited strength model. First, this paradigm can only provide evidence for the existence of the depletion effect, a phenomenon that must exist if the resource depletion hypothesis is correct, but one that is not sufficient to prove that the hypothesis is correct. At no point in the sequential task paradigm is self-regulatory strength measured, and therefore, rejection of the null hypothesis (i.e., demonstrating the depletion effect) *is not evidence for the depletion of some resource*. Currently, there is no compelling reason to think that an actual resource exists, or that some alternative explanation (for example, that the depletion effect is purely caused by diminished motivation), is less plausible. This lack of direct evidence is sometimes ignored by proponents of the limited strength model, and when referring to evidence for the depletion effect derived from the sequential task paradigm, appropriately metaphorical language is occasionally abandoned in favor of statements such as "Once again, the findings showed that acts of self-control depleted an important resource" (Baumeister, Muraven, & Tice, 2000; p. 133) or "Empirical tests have shown that self-regulatory resources underlie a wide range of behaviors across a variety of domains… (Vohs & Faber, 2007; p. 538). This dissertation focuses exclusively on the use of the sequential task paradigm, and so it offers no new evidence regarding the existence of a resource. As noted by Hagger et al., the results of a meta-analysis of the sequential task paradigm can only address whether or not the depletion effect is robust, which constitutes a necessary condition for the resource depletion hypothesis, but is not sufficient to demonstrate that self-regulatory strength exists.

A second consequence of the heavy reliance on the sequential task paradigm is that there is very little methodological consistency in the tasks that have been used to induce or measure the depletion effect. In fact, although there appears to be a core set of tasks that researchers believe to require use of the active self (e.g., Hagger et al. identified ten frequently used manipulation and outcome tasks), the literature on the depletion effect is notable for an overall lack of consistent methodology, and is probably best described as a collection of conceptual replications and theoretical extensions. It seems likely that the this methodological heterogeneity is due in part to the fact that the sequential task paradigm was explicitly designed to readily incorporate a wide range of tasks (Baumeister, et al., 1998), hence the large number of conceptual replications; but the diversity also appears to be a product of attempts to use the sequential task paradigm to test hypotheses about whether the performance of other behaviors require use of the active self (i.e., theoretical extensions). For example, Gailliot, Schmeichel, and Baumeister (2006) reported several studies in which they investigated whether greater levels of self-regulation were related to better alleviation of thoughts and feelings about death. These studies were based on the assumption that "…people periodically confront cues or thoughts that remind them of death and that they therefore exercise self-regulation to prevent these thoughts from lingering in conscious awareness and escalating into greater anxiety" (p. 52). On this assumption, Gailliot et al. reported two experiments using the sequential task paradigm to examine whether the use of self-regulatory resources resulted in greater preoccupation with death. The authors observed that (a) greater use of self-regulation on an initial task seemed to result in a greater accessibility of death thoughts, and that (b) writing about death, rather than about uncertainty, seemed

to cause participants to perform worse on analytical reasoning problems. Gailliot et al. described their results as evidence that "self-regulation is a key intrapsychic mechanism for alleviating troublesome thoughts and feelings about mortality" (p. 49), a conclusion at which they were able to arrive by assuming that the depletion effect is real and then showing that thoughts of death (either manipulated or measured) functioned in a way similar to behaviors that require the active self.

Producing a meta-analytic test of the depletion effect requires that the analyst make decisions about which tasks, out of the multitude of possible tasks, should be considered to draw on the active self. Therefore, a critical step in this dissertation was to develop a way to account for the fact that the literature on the depletion effect is made up of primarily (a) experiments designed to test the same effect via very different methodologies (i.e., conceptual replications), and (b) experiments designed under the assumption that the depletion effect is real (i.e., theoretical extensions). The way in which I approached this issue is discussed at length below.

Current evidence for the depletion effect

In 2010, Hagger, Wood, Stiff, and Chatzisarantis published a meta-analysis designed to assess the current evidence for the existence of the depletion effect. The term *meta-analysis* refers to "the statistical analysis of a large collection of analysis results from individual studies for purposes of integrating the findings" (p. 3; Glass, 1976). This process is believed to provide the most definitive and statistically valid answer to a research question by quantitatively summarizing the results of many different experiments on the same hypothesis. Applied to the depletion effect as tested via the

sequential task paradigm, such an analysis should provide the best evidence for whether the depletion effect is real and robust.

Hagger et al. (2010) collected and meta-analyzed 83 published studies on the topic of depletion, which amounted to 198 independent experiments that used the sequential task paradigm, conducted on 10,782 participants between 1998 and April 1st, 2009. The overall average decrement in performance due to the depletion effect was estimated as $d = 0.62$ with significant between-study heterogeneity, $Q = 301.79$, $p < .001$, $I^2 = 34.72$, meaning that, on average, performance of participants in the experimental groups was 0.62 standard deviations different from that of participants in the control groups in the direction predicted by the limited strength model, but that there was also evidence for a non-zero amount of unexplained variance between effect sizes. The between-study heterogeneity suggests that the overall effect is an average of effect sizes that are actually estimates of a variety of effects—that is, the effect sizes that have been aggregated together appear to be drawn from different populations of effect sizes. Generally, between-study heterogeneity is thought to be due to observable study characteristics, such as differences in experimental design, differences in the demographic characteristics of the sample used, or differences in methodology between experiments (i.e., moderators of the effect). To address the significant between-study heterogeneity, Hagger et al. created subsets of effects sizes based on a more fine-grained assessment of the methods of individual studies. For example, all experiments that used a manipulation task that seemed to require the control of impulses were meta-analyzed together. Within all subset meta-analyses, the estimates of the overall effects were found to remain positive and different from zero, despite some slight variation in magnitude.

Overall, the results from Hagger et al (2010) suggest a robust effect for depletion, and were described as "…demonstrating that the ego-depletion effect exists, its associated confidence intervals do not include trivial values, and it is generalizable across spheres of self-control" (p. 515). It would seem that the results of the synthesis by Hagger et al. are good evidence that the depletion effect exists. However, there are three major reasons to think that the methods used by Hagger et al. constitute a test that is biased toward confirming the existence of the depletion effect.

Reason one: The inclusion of any instance of the sequential task paradigm, regardless of the tasks used. The first issue that may have resulted in a biased estimate by Hagger et al. is how the authors approached the methodological heterogeneity present in tests of the ego-depletion effect. Between-study heterogeneity can be sorted into two classes: (a) methodological heterogeneity, such as differences in study design, study location, and participant characteristics, and (b) statistical heterogeneity, which is present when the studies in a meta-analytic sample are derived from different underlying effects (Higgins & Thompson, 2002). Statistical heterogeneity may be apparent if the variation between effect sizes in a set of studies is greater than the variation one can expect from sampling error alone (a number of procedures exist for quantifying statistical heterogeneity; Rücker, Schwarzer, Carpenter, & Schumacher, 2008). Importantly, methodological heterogeneity may or may not be responsible for statistical heterogeneity (Higgins & Thompson, 2002), and it should be made clear that a test for statistical heterogeneity is not a test for methodological heterogeneity. Therefore, determining whether an experiment is methodologically similar enough to a given set of experiments

is not a question that can be resolved purely through a quantitative test, but involves qualitative judgments by the analyst.

This last point is particularly important for a synthesis of the literature on the depletion effect because, as described above, the literature is characterized by a large degree of methodological heterogeneity due to the presence of conceptual replications and theoretical extensions. These conceptual replications are difficult to integrate into a meta-analysis because there is usually insufficient evidence for construct validity for tasks used in the sequential task paradigm: Specifically, for many conceptual replications, it is rare that efforts are made to establish that any one task does indeed require use of the active self, and researchers tend to settle for face validity instead. Including these experiments in a meta-analysis based on the arguments for face validity made by the original authors is not an ideal approach, as there is no objective way to determine whether a task actually draws on the active self. In contrast, deciding whether to include theoretical extensions in a meta-analysis of the depletion effect is an easier choice, as a theoretical extension is not, strictly speaking, a test of the depletion effect, and therefore cannot provide information about whether the depletion effect is real.

Including a set of studies in a meta-analysis implies that one believes that the studies all represent estimates of a single phenomenon (Hedges, 2009), so including experiments that may not actually test the depletion effect, either because the tasks involved do not have established construct validity or because the experiment was not designed to test the effect, will likely lead to an inaccurate estimate. Hagger et al. elided this problem by including both conceptual replications and theoretical extensions in their

meta-analysis, as long as the experiments represented instances of the sequential task paradigm.

Reason two: Ignoring contradictions and ambiguous results in the literature. The second issue that likely increased the chances of Hagger et al.'s results confirming the existence of the depletion effect is the way in which various contradictions and inconsistencies in the ego depletion literature were handled. In the literature on the limited strength model, there are a variety of contradictions and ambiguous results (e.g., different authors claiming that seemingly opposite results are evidence that the depletion effect exists). Hagger et al. coded results as confirming the depletion effect based on whether the authors interpreted their results as such, even when the authors of different articles had made contradictory theoretical interpretations of essentially identical empirical phenomena. For example, Janssen, Fennis, Pruyn, and Vohs (2008) conducted an experiment in which all participants were first asked to watch a silent video of a woman being interviewed by an off-screen interviewer. During this video, words appeared in one corner of the screen. Participants in the experimental group were told to watch only the woman and to ignore the words, whereas participants in the control condition were told to watch the video as they would watch any other video. Following this manipulation, participants were given the chance to donate some of the money they had been paid for participating in the experiment to a charity that was working toward developing educational projects in third world countries. For half of the participants, the charity was described as well-known, renowned, and experienced; and for the other half of participants, the charity was described as unknown and as having only recently begun relief work. The prediction of Janssen et al. (2008) was that depleted participants would

rely on heuristic-based decision-making and only give to the charity that seemed the most authoritative (i.e., the well-known, experienced charity). As predicted, the authors observed a significant interaction effect, for which a decomposition of the simple effects showed that participants who had controlled their attention during the video donated a significantly larger sum of money compared to those who had not controlled their attention, but only when they had been told the charity was renowned, $d = 0.94$. Hagger et al. coded this effect as $d = 0.94$, as depletion was not hypothesized to exist amongst the participants who were told the charity was new.

DeWall, Baumeister, Gailliot, and Maner (2008) reported the results of a similar experiment. Participants in DeWall et al.'s experiment experienced the same manipulation as those in Janssen et al. (2008)'s experiment. Following the video watching task, DeWall et al. (2008) had participants listen to a recording of a fake radio broadcast that recounted the story of a woman who had recently experienced a tragedy and was attempting to raise money for herself and her family. Following this, participants were told the experiment had ended, but that if they were interested, the professor in charge of the research was organizing a volunteer effort to help the woman they had learned about during the session. The number of hours participants volunteered to help was the dependent variable. Given what was found by Janssen et al., one might expect that depletion (caused by the same attention-control task) would result in more hours volunteered; however, the opposite was found—participants who controlled their attention during the initial task volunteered fewer hours than did those participants who had not controlled their attention, $d = 0.96$. Hagger et al. coded the effect size for this study as $d = 0.96$. The explanation—and the theoretical extension of the limited strength

model that had initially inspired DeWall et al. (2008) to conduct the experiment—was that self-regulation is necessary to overcome selfish motives in favor of prosocial ones—an explanation that should have led one to expect a negative main effect of depletion in the Janssen et al. (2008) paper, rather than the increase in donations to renowned charities by depleted participants that they observed. The results from Janssen et al. (2008) and DeWall et al. (2008) exemplify the inherent difficulties in synthesizing results from tests of the limited strength model. From the perspective of the authors (and of Hagger et al.), both studies support the depletion hypothesis. But there is an obvious argument to be made that the two sets of results are contradictory and that either (a) one of the two should have been coded as negative; or (b) the limited strength model does not lead to clear predictions regarding charitable donations and volunteerism and that the results of these two papers perhaps should not have been included in a meta-analysis designed to evaluate the depletion effect.

A similar situation arises when outcome tasks, such as the Stroop task, furnish more than one dependent variable. The Stroop task usually involves presenting a participant with a color-word (e.g., *blue*) that is displayed in a color that is either congruent with the word (e.g., blue-colored font) or incongruent with the word (e.g., red-colored font). The participant is then asked to avoid reading the word and instead respond by identifying the color of the font in which the word is presented. It is argued that the Stroop task requires self-regulation in that, to be successful, one must override a habitual response of simply reading the target word. Critically, the standard Stroop task usually furnishes four outcome measures: Reaction time, accuracy, and the difference in performance (either reaction time or accuracy) between incongruent trials and congruent

trials. One might expect that depletion should be evident in all of these variables, but this is not always the case. For example, the seventh experiment reported in Gailliot et al. (2007) exposed participants to the attention video described above followed by the Stroop task (80 trials, all of which were incongruent). The depletion effect was evident in accuracy scores, but not in reaction time. In contrast, Inzlictht and Gutsell (2007) had participants watch emotionally disturbing videos, where half of the participants had been told to suppress all internal and external reactions to the videos and half had been told to watch normally. Following the videos, participants completed the Stroop task (864 trials, 288 of which were incongruent). Inzlicht and Gutsell found no differences between groups for reaction time or for accuracy, but they did find a significant interaction in which the depleted participants performed slower on incongruent trials only. Hagger et al. coded both of these experiments as demonstrating the depletion effect, despite the fact that the two experiments produced contradictory results. A better solution would have been to aggregate across the variables produced by a single task, so that each experiment provided some form of average between accuracy and reaction time scores. Of course, doing so would have resulted in each study providing considerably weaker evidence in support of the ego depletion phenomenon. Instead, Hagger et al.'s approach likely biased their meta-analysis in favor of confirming the depletion effect.

Reason three: Potential for publication bias. The third and most clear-cut issue limiting the usefulness of Hagger et al.'s (2010) otherwise impressive work is that they included only published data in their meta-analysis, a methodological choice that makes a meta-analysis invalid to the extent that published studies are unrepresentative of the entire population of studies (Rothstein, Sutton, & Borenstein, 2005; Sutton, 2009).

Publication bias refers to a collection of biases that influence the dissemination of results. For example, researchers may conduct many tests of a single hypothesis, but may only publish a portion of such tests, perhaps the ones that are believed to be most interesting, accurate, or supportive of the authors' favored hypothesis. Often, the perception of interestingness and accuracy of an empirical finding is a function of its statistical significance. There is evidence that statistically significant findings are more likely to be published (Chan et al., 2004), and that such findings are published more quickly than are nonsignificant findings (Stern & Simes, 1997). To the extent that experiments using the sequential task paradigm were more likely to be published if they reached statistical significance, then the sample of effect sizes that were synthesized by Hagger et al., and the resulting statistical summary, are biased in favor of confirming the existence of the depletion effect.

Although Hagger et al. were certainly not ignorant of publication bias, they did not take the most important first step to minimize it, which is to search for unpublished results to include in the overall meta-analysis. Instead, they addressed publication bias post-hoc using a variant of a method known as the failsafe *N* (Rosenberg, 2005). The failsafe *N* estimates the number of effect sizes that would need to be included in a meta-analytic sample to reduce the overall effect size to statistical nonsignificance, given that those missing effect sizes were equal to zero on average. Hagger et al. found that, for the overall effect, 50,445 null effect sizes would need to exist to bring their estimate down to nonsignificance. This number exceeds a cutoff above which it is thought to be unlikely that so many such effect sizes exist, published or otherwise (Rosenthal, 1979).

Failsafe methodology is questionable. In fact, such methods have been called "nothing more than a crude guide" and are thought to "lead to unjustified complacency about publication bias" (p. 443; Sutton, 2009). The variant of the failsafe *N* method proposed by Rosenberg (2005) and used by Hagger et al. (2010) is superior to other failsafe *N* methods (e.g., Rosenthal, 1979), but it still has serious limitations. For example, as with other failsafe methods, Rosenberg's (2005) failsafe *N* is based on the assumption that the studies that are missing are of equivalent sample size to the studies in the meta-analysis, but to the extent that missing studies are smaller or larger, then the failsafe N will be an under- or over-estimate, respectively. Additionally, the use of Rosenberg's (2005) method assumes that all missing studies have an average effect of zero (or, for a random-effects meta-analysis, that all missing studies have effect sizes of *exactly* zero). It is easily possible for missing studies to have an average effect size of less than zero, and this would presumably be the case if nonsignificant findings were being selectively left out. If these unrealistic assumptions are incorrect, the failsafe method will produce an inaccurate result, and because Hagger et al. did not include unpublished studies, it is impossible to glean any information about whether the assumptions of the failsafe method were fulfilled. The addition of other methods to identify and correct for publication bias is a necessity.

The current studies

For the reasons outlined above, critical readers should not feel satisfied with the validity of the estimate of the depletion effect provided by Hagger et al. This is unfortunate because greater certainty about whether the depletion effect exists will allow future work to rest on firmer ground, and the large and varied group of researchers

interested in self-regulation could build from these foundations. The goal of this dissertation was to provide a meta-analytic estimate of the depletion effect in which even skeptical readers can have confidence. To this end, I conducted two studies. In Study One, to illustrate the potential influence of publication bias on Hagger et al.'s conclusions, I re-analyzed Hagger et al.'s original dataset of 198 studies using more appropriate methods for assessing and correcting for publication bias. In Study Two, I conducted an updated meta-analysis by (a) including both published and unpublished experimental work on the sequential task paradigm, (b) bringing the database up to date by including experimental work that has been completed since the publication of Hagger et al. (2010), and (c) refining the scope of the quantitative review by focusing only on experiments that involve tasks that are frequently used by researchers testing the limited strength model.

Chapter Two: Study One Method

In Study 1, I re-analyzed the data in Hagger et al. (2010) using a set of statistical techniques that are designed to assess and correct for publication bias. Upon request, Martin Hagger provided me with the coded effect sizes, given as Cohen's *d*, for each experiment and the *n*s for the depletion and control groups. Notably, Hagger et al. modified three "outlier" effect sizes, which might potentially have obscured funnel plot asymmetry. Therefore, I used the original, unmodified effect sizes. To complete my analysis, I needed to also calculate the variance associated with each effect size: The standardized difference between two population means (μ_1 and μ_2) with equal population standard deviations ($\sigma_1 = \sigma_1 = \sigma$) is δ, where

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}.$$

An estimate of the standardized difference, *d*, can be calculated from sample means, Y_1 and Y_2, with sample sizes, n_1 and n_2, and sample standard deviations, S_1 and S_2, as

$$d = \frac{Y_1 - Y_2}{S_{within}},$$

where S_{within} is the pooled standard deviation and is calculated as

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

The variance of *d*, v_d, represents the combined uncertainty of the estimates for the mean difference and for the pooled standard deviation and is calculated as

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}.$$

Using the data provided by Martin Hagger, I calculated v_d for each effect size. Thus, the dataset included 198 observations, each with an effect size, a corresponding variance, and a sample size for both the depletion and control groups.

Statistical analysis

Unless otherwise specified, analyses were conducted using R (version 2.15; R Development Core Team, 2011) with the metafor package (Viechtbauer, 2010).

Random-effects model. To begin my re-analysis of Hagger et al.'s dataset, it was first necessary to re-estimate the basic meta-analytic models. Meta-analysis is generally conducted using either a fixed-effect (FE) model or a random-effects (RE) model. Borenstein, Hedges, Higgins, and Rothstein (2010) have recommended that a FE model be used if the following two conditions are met: (a) all studies are likely to be functionally identical, and (b) the goal of the meta-analysis is to compute an effect size that is not necessarily generalizable beyond the population included in the analysis. The first condition is often implausible, and in the current study, most likely incorrect given the obvious differences in tasks that are used as manipulations and dependent measures. The second condition, which is called *conditional inference*, suggests that a FE approach would be more limiting than desirable: The goal of the current project was to derive inferences about the hypothetical set of all experiments that have been run, could have been run, or will be run to test the depletion effect—that is, based on the current study, one should be able to draw conclusions about the larger, hypothetical set of experiments from which the current studies are (assumed to be) a random sample (this is known as *unconditional inference*). For these reasons, I used RE models, based on methods

described by Raudenbush (2009), which decompose the variation among effect size estimates into two different "levels."

If one considers the sample of studies to be meta-analyzed as a random sample of a larger set of hypothetical studies in the way described above, then the sampling procedure that resulted in the observed effect sizes that make up the current sample can be conceived of as having two stages. Each stage of sampling is associated with a corresponding variance component. The first stage represents the process of sampling subjects into an individual study to estimate the true effect size dependent on the specific study characteristics that produced the estimate. The variance component that arises from that process is called *estimation variance*, or *sampling error*. This is represented at level one of the model, and it is equivalent to the FE model.

The second stage represents the process of randomly sampling studies into the current sample of studies to be meta-analyzed. Each study produces an observed independent effect size, which can be thought of as an estimate of the true effect size that is unique as a function of observed study characteristics. The variance component that arises from the second sampling process is called *random effects variance*, and represents unobserved sources of variation or variation not represented by factors included in the model (i.e., variation not due to sampling error or to any observed study characteristics that are explicitly included in the model). This is represented at level two of the model.

More specifically, for k observed effect sizes, $i = 1, \ldots, k$, level one is decomposed as follows:

$$T_i = \theta_i + e_i,$$

$$e_i \sim N(0, v_i),$$

where, for the i^{th} study, T_i is the observed effect size that is an estimate of the unknown true effect size, θ_i, and differs from θ_i by sampling error, e_i (which is normally distributed about a mean of zero with an observed variance of v_i). Level two is decomposed as follows:

$$\theta_i = \mu + u_i,$$

$$u_i \sim N(0, \tau^2),$$

where μ is the mean of the unknown true effect sizes that are estimated by the observed effect sizes, and u_i is a random effect for the i^{th} study representing the deviation of the i^{th} study's true effect size from the mean of the true effects (in other words, $u_i = \mu - \theta_i$). u_i is normally distributed about a mean of zero with a variance of τ^2. τ^2 is the total amount of statistical heterogeneity among the true effects (i.e., the degree to which the true effects represent different phenomena), and when it is zero, this suggests homogeneity—that is, $\theta_1 = \ldots = \theta_k \equiv \theta$ and $\mu = \theta$. Additionally, when τ^2 is zero, the FE and RE models are the same: A τ^2 of zero means that $u_i = 0$, so $\theta_i = \mu$, and each T_i differs from θ_i only by e_i (this signifies that all variation in the true effect sizes is due to sampling error, which is the defining feature of the FE model). The RE model can be used to estimate the size of μ, that is, the average of the true effects within the larger, hypothetical set of experiments from which the current set of experiments are thought to be a random sample. This estimation allows for unconditional inference, and can be interpreted as the estimate of the overall effect.

RE models are fitted in a two-step approach (Raudenbush, 2009). For the first step, τ^2 is estimated using one of a variety of estimators. For this dissertation, I used restricted maximum-likelihood estimation (Viechtbauer, 2005; Viechtbauer, 2010). Based

on the estimate of τ^2 derived from the first step, the null hypothesis that all heterogeneity has been accounted for by the model (i.e., $\tau^2 = 0$) was tested using Cochran's *Q*-test (Hedges & Olkin, 1985) which is described in more detail below. For the second step, point estimates are obtained using weighted least squares estimation. The weights used in this estimation are the reciprocal of the estimated variance for each study, given as $\hat{v}_i^*$. In other words, the estimated variance of the i^{th} effect size is $\hat{v}_i^* = v_i + \hat{\tau}^2$, where $\hat{\tau}^2$ is the estimate of τ^2 from the first step, and the weight used for the i^{th} effect size is $1/\hat{v}_i^*$.

For example, the estimate for average overall effect, μ, written as d^*, would be given by

$$d^* = \frac{\sum_{i=1}^{k} \hat{v}_i^{*\,-1} T_i}{\sum_{i=1}^{k} \hat{v}_i^{*\,-1}}.$$

The standard error for this estimate can be calculated as the square root of the estimated variance, or $\hat{Var}\,(d^*)$, where

$$\hat{Var}\,(d^*) = \frac{1}{\sum_{i=1}^{k} v_i^{*-1}}.$$

Using this information, the null hypothesis that the point estimate is zero, H0: $d^* = 0$, can be tested with a *t* distribution[3] (with *k-p* degrees of freedom at an $\alpha = 0.05$ level), so that the test-statistic is

$$t = \frac{d^*}{\sqrt{\hat{Var}\,(d^*)}}.$$

As mentioned, the null hypothesis H0: $\tau^2 = 0$ is tested using Cochran's *Q*-test (Hedges & Olkin, 1985). Q_0 is given as

$$Q_0 = \sum_{i=1}^{k} v_i^{-1} (T_i - d^*)^2 .$$

Q_0 follows a chi-squared distribution with degrees of freedom of k - 1. The null hypothesis H0: $\tau^2 = 0$ can be tested by comparing the obtained Q_0 to the appropriate chi-squared distribution at an $\alpha = 0.05$ level. If the Q-test is significant, this suggests non-zero heterogeneity.

Publication Bias and Small Study Effects. The following methods for assessing and correcting for potential publication bias were used.

Assessment methods based on estimates of statistical power. Statistical power is defined as the probability that the null hypothesis will be correctly rejected at a given α level. Power is represented as $1 - \beta_i$, where β_i is the probability of incorrectly accepting the null hypothesis, or the type II error, for the i^{th} test. Power is determined by the magnitude of the true effect size for the phenomenon being examined with the i^{th} test, the alpha level of the i^{th} test, and the sample size used to conduct the i^{th} test.

A recent method for assessing publication bias based on statistical power has been proposed. This test takes two forms: The first version, proposed by Ioannidis and Trikalinos (2007a), evaluates whether there is an excess of statistically significant results in a published literature by comparing the observed number of statistically significant findings to the expected number. The second version of this test, proposed by Schimmack (2012), is called the incredibility index (IC-index), and is simply the inverse of the exploratory test of Ioannidis and Trikalinos (2007a)—that is, it is the probability that the number of non-significant findings were observed given the expected number of non-significant findings (literally 1 minus the p value from the exploratory test). The IC-index, therefore, provides information about both the improbability that a set of studies includes as many non-significant findings as it does and the probability that a set of

studies yielded more non-significant studies than are reported. These methods make no assumptions about the actual cause of the discrepancies between observed and expected findings, which, in reality, may be due to such sources as publication bias, flawed execution or reporting of statistical tests in the primary literature, or randomness.

Because they are based on statistical power, both methods require an estimate of a true underlying effect size. Given a set of studies, power can be calculated in two ways. First, it is possible to obtain an estimate of the true effect size using information from the entire set of studies (as in a meta-analysis, for example), and to then calculate power for each study given the estimated true effect (or given a range of possible true effects, such as the upper and lower limits of the 95% confidence interval surrounding a meta-analytic estimate of an effect size). Second, it is possible to use the effect size estimates from each study to calculate post-hoc power for each study independently. In both cases, the power estimates can be averaged to give the expected proportion of statistically significant results in a set of studies, *E*. The observed number of studies with statistically significant results, *O,* is then compared to *E* using a binomial test with the null hypothesis that *O* is likely given *E*. A chi-squared test is also possible, but is ill-advised when *n* is small (Ioannidis & Trikalinos, 2007a). The significance of this binomial test reflects the probability that the number of significant studies exceeds chance levels, or, when inverted (i.e., the IC-index), the probability that more non-significant findings exist than are observed. For the current study, the test was described in terms of the IC-index.

A test for significance for the binomial test, $p < 0.10$ (which corresponds to values of the IC-index that are greater than 0.90), was suggested by Ioannidis and Trikalinos (2007a). However, following Schimmack (2012), I use the IC-index less as a hard line for

making dichotomous decisions about the existence of bias, and more as a descriptive statistic. The IC-index was based on power calculated using (a) the estimated true effect, d^*, from the RE meta-analysis model, (b) with the upper and lower limits of the 95% confidence interval for the estimate of d^*, and (c) the effect sizes of each individual study. For the these methods, power calculations were computed by treating each effect size as an estimate derived from an independent samples *t*-test.

Assessment and adjustment methods based on the funnel plot. A funnel plot is a specific type of scatter plot in which effect size is plotted on the x-axis against some measure of statistical precision (i.e., the inverse of the standard error of an effect size estimate) along the y-axis. When plotted in this way, and when the set of effect sizes are both unbiased and representative of a single true effect, the data points in the plot look like an inverted, symmetrical funnel because the effect sizes that best represent the true effect size tend to cluster around that value while smaller studies produce more variable estimates of the true effect size and tend to fall to the right and left on the true effect size (Light & Pillemer, 1984). When the sample of effect sizes is biased, this can be apparent in the shape of the funnel plot. For example, if selection exists such that non-significant findings are not published, and therefore not included in the sample of studies being meta-analyzed, then there will be an apparent gap in the funnel shape around zero on the x-axis. This gap results in asymmetry of the funnel shape of data points in the funnel plot.

The meta-analyst may visually inspect a funnel plot for evidence of bias (i.e., asymmetry), but this method is questionable (Terrin, Schmid, & Lau, 2005). Therefore, various techniques have been proposed (see Moreno et al., 2009) to statistically model the relationship between effect size and some estimate of precision. For this dissertation,

the relationship was modeled in three ways. The first two approaches are weighted least squares (WLS) regression models (Stanley & Doucouliagos, 2011; Moreno et al., 2009), and the third approach is the non-parametric trim and fill method (Duval & Tweedie, 2000a; 2000b). For the regression-based methods, the regression weights are given as the inverse of variances for the individual effect sizes, v_i^{-1}. The first method was suggested by Stanley and Doucouliagos (2011) (see also Stanley, 2008), and it involves regressing effect size on standard error (this model is the same as the one proposed by Egger, Davey Smith, Schneider, and Minder [1997] to assess for funnel plot asymmetry):

$$T_i = b_0 + b_1\sqrt{v_i} + e_i \, .$$

Use of this model to assess and correct for small study effects is sometimes called "precision-effect testing," or PET (Stanley & Doucouliagos, 2011). The second used was recommended by Moreno et al. (2009), and it involves regressing effect size on variance:

$$T_i = b_0 + b_1 v_i + e_i \, .$$

This model is a modified form of PET, and its use is sometimes referred to as "precision effect estimate with standard error," or PEESE (Stanley & Doucouliagos, 2011).

For PET, the statistical significance of b_1 is often interpreted as a test for whether funnel plot asymmetry exists (Egger et al., 1997). For both PET and PEESE, b_0 can be interpreted as the estimated effect size of a hypothetical study with an infinitely large sample size (i.e., zero sampling error variance). Use of the intercepts from both PET and PEESE has been suggested as an estimate of the true effect size uninfluenced by small study effects, such as publication bias (Stanley & Doucouliagos, 2011; Moreno et al., 2009). Stanley and Doucouliagos (2011) suggest that PET and PEESE are most effective when used in conjunction, because there is evidence that PEESE under-corrects bias

when the true underlying effect is no different from zero. When used together, one should only turn to results from PEESE if b_0 from PET is significant. This issue has not been firmly resolved, though Stanley and Doucouliagos (2011)'s argument is persuasive. In the interest of completeness, I apply and report the results from both models.

Another method for assessing and correcting for publication bias that is based on funnel plot asymmetry is known as the trim and fill (Duval & Tweedie, 2000a; 2000b). Trim and fill is an iterative nonparametric test that both estimates that number of missing studies and corrects the estimate of the overall effect size for funnel plot asymmetry. First, each effect size has the overall effect size estimate subtracted from it. Next, each effect size is assigned a rank based on its absolute value (where higher ranks represent greater deviation from the overall effect size). Ranks are given a negative sign if they represent effect sizes that were less than the overall effect size (e.g., in a case with seven effect sizes, the second and fourth smallest of which have an absolute value of less than the average effect size, the ranks would be given as -4, -2, 1, 3, 5, 6, 7). At this point, an estimator of the number of missing studies is used. There are two primary estimators, called R_0 and L_0, which can be used for trim and fill; however, the two estimators tend to give similar results and so, following Moreno et al. (2009), I will only use the R_0 estimator. R_0 is defined as

$$R_o = \gamma^* - 1,$$

where γ* is the longest consecutive run of positive ranks (so in the example above, γ* = 3). Once R_0 is determined, the R_0 studies with the largest effect sizes are deleted ("trimmed"). This process is done iteratively until the estimate for the number of missing studies converges, after which hypothetical missing studies are added ("filled") and the

overall effect size recalculated. The estimated number of missing studies gives the analyst a sense of the degree of funnel plot asymmetry and the overall effect size calculated on the filled data gives an estimate of the effect size adjusted for that asymmetry.

Statistical heterogeneity and methods for assessing and correcting for publication bias. Concerns have been raised in the literature about the application of each of the above techniques in the presence of moderate to extreme statistical heterogeneity (e.g., I^2 > 50%). The theoretical basis for these concerns is that statistical heterogeneity suggests that multiple true underlying effects are being measured by studies in the meta-analytic sample, and thus, the assumption upon which the above techniques are based cannot be said to hold (Terrin & Schmid, Lau, & Olkin, 2003; Ioannidis & Trikalinos, 2007a; 2007b).

Substantial statistical heterogeneity seems to be particularly problematic for the IC-index and the trim and fill—which are both methods designed specifically for assessing missingness due to publication bias (Schimmack, 2012; Terrin et al., 2003 Ioannidis & Trikalinos, 2007a; 2007b); however, statistical heterogeneity is less of a concern for the more general regression-based methods, PET and PEESE. If one conceptualizes funnel plot asymmetry as a special case of statistical heterogeneity (referred to as small-study effects; Rücker, Carpenter, & Schwarzer, 2011), then the logical next step is to explain this heterogeneity by examining possible meta-analytic moderators. A common tool for explaining heterogeneity is meta-regression, which allows the analyst to statistically control for the effect of some moderator on the underlying effect of interest (Rücker, et al., 2011). Importantly, this is exactly what PET

and PEESE are designed to do: Statistically control for the influence of small-study effects on the underlying effect of interest. From this perspective, heterogeneity is therefore not a problem for the application of PET and PEESE, but a necessary precondition (Rücker, et al., 2011).

Based on simulation studies, PET and PEESE do appear to become relatively more inaccurate in the face of extreme statistical heterogeneity (e.g., Stanley & Doucouliagos, 2011; Moreno et al., 2009). Therefore, it is important to keep the amount of statistical heterogeneity in mind when interpreting results for these methods. However, these regression-based methods have consistently outperformed the trim and fill in terms of reducing the inflation of effect size estimation due to publication bias, and they have been shown to be quite accurate (Stanley, 2008; Moreno, et al., 2009; Rücker, et al., 2011; Stanley & Doucouliagos, 2011). Additionally, when interpreting results from PET and PEESE, one should also bear in mind that these methods do not necessarily measure publication bias. Publication bias, as described above, is one type of small-study effect, but others are possible—for example, a particularly potent experimental manipulation may be time consuming and expensive to collect, and thus, only smaller samples can be collected when using it. In such hypothetical smaller samples, the effect will be more potent, and thus, an association will exist between effect size and sample size. Regardless of the nature of the small-study effect, PET and PEESE produce an estimate of the underlying effect when the influence of the small-study effect is held at zero, but it should be recognized that publication bias is not the only possible small-study effect.

Chapter Three: Study One Results and Discussion

The RE meta-analysis model, the IC-index, the trim and fill, and PET and PEESE were calculated using Hagger et al.'s original dataset. A contour-enhanced funnel plot (Peters, Sutton, Jones, Abrams, & Rushton, 2008) is displayed in Figure 1. As recommended, a one-tailed test was used in the binomial test (Ioannidis & Trikalinos, 2008) and $p < 0.10$ was used as the cutoff for tests of funnel plot asymmetry (Egger et al., 1997).

First, based on the RE model, the average overall effect was d* = 0.68, 95% confidence interval (0.63, 0.74). The Q test was statistically significant, Q = 320.68, p < 0.001. τ^2 was estimated as 0.05, and I^2 = 38.6%, suggesting a moderate amount of between-study heterogeneity. These results are similar to what was reported by Hagger et al.

Second, the IC-index indicated that the observed number of significant findings exceeded the expected number, regardless of the effect size estimate used to calculate power: The estimate from the individual experiments (IC-index = 0.999), the estimate from the RE model (IC-index = 0.999), or the estimates from the upper (IC-index = 0.995) and lower (IC-index = 0.999) limits of the CI given by the RE model).

Third, the trim and fill method required that the sample be increased by 73 experiments, or 37%, to achieve funnel plot symmetry, and the estimate of the overall effect was reduced by 26% to $d^* = 0.50$. Examination of the contour-enhanced funnel plot (Figure 1) suggests that asymmetry is mainly due to a lack of data points in the area of statistical nonsignificance, which is consistent with the influence of publication bias, rather than some other small-study effect.

Third, according to the coefficients in the regression models, there was clear evidence for funnel plot asymmetry: $b_1 = 2.72$ ($p < 0.001$) and $b_1 = 4.74$ ($p < 0.001$) for PET and PEESE, respectively. Furthermore, results from applying PET and PEESE strongly suggest that the true underlying effect for the overall sample and each of the subsamples is not different from zero: $b_0 = -0.10$ ($p = 0.11$) and $b_1 = 0.25$ ($p < 0.001$) for PET and PEESE, respectively. As mentioned above, since the estimate of b_0 from PET was nonsignificant, it likely a more accurate estimate of the true underlying effect than b_0 from PEESE, which will tend to overestimate the underlying effect when it is zero (Stanley & Doucouliagos, 2011).

Based on these results, it seems very likely that publication bias led to an overestimation of the depletion effect by Hagger et al. Indeed, it appears that the influence of publication bias is so strong, that it may be the sole explanation for the apparent existence of the depletion effect.

Although informative, Study One only served to illustrate my point that publication bias could have led to an overestimation of the depletion effect by Hagger et al. To deal with the two additional reasons for suspecting Hagger et al.'s estimates, as well as to bring the dataset up to date, I conducted Study Two.

Chapter 4: Study Two Method

As mentioned above, an important feature of Study Two was that the scope of the meta-analytic effort was refined to help deal with the issues created by the wide range of methods and interpretations that are present in the literature on the depletion effect. I only included data from experiments in which both the task used as the manipulation and the task used as the outcome measure have been frequently used in the sequential task paradigm literature. The goal here was not to reduce the likelihood of confirming the existence of the depletion effect, or to exclude any particular studies: Rather, the goal was to avoid the need for subjective judgment regarding both the validity of tasks used in the sequential task paradigm and the interpretation of results. There are four reasons for focusing only on experiments that make use of frequently used measures of depletion:

First, focusing on only those tasks that are most frequently used does not limit the ability of the meta-analysis to test for the depletion effect. According to the limited strength model, any and all combinations of tasks that involve the active self should create depletion, so showing that one or a number of combinations of these tasks do not create ego depletion will be informative.

Second, by definition, refining the range of analyses to only the most frequently used tasks ensures that this meta-analysis will focus on the tasks for which researchers have shown the most interest.

Third, as mentioned above, instances of the sequential task paradigm are often conceptual replications that make use of tasks on the basis of face validity, or theoretical extensions that use the sequential task paradigm to examine other predictions. It may be that frequent use denotes a confidence on the part of researchers that these tasks are the

most valid ways to manipulate use of the active self, so using the approach of only examining a small set of frequently used tasks should ensure that only those tasks thought to have the highest construct validity are included, and that theoretical extensions that should not be considered tests of the ego depletion hypothesis are excluded. Note that this approach does not confirm that these tasks actually have construct validity, but rather, that researchers interested in the depletion effect act as if they believe that they do.

Finally, both methodological and statistical heterogeneity make the application and interpretation of methods to correct and assess for publication bias difficult to interpret (e.g., Terrin, Schmid, Lau, & Olkin, 2003; Moreno et al., 2009). Because one of the main goals of the current study is to assess and correct for any potential publication bias, it was necessary to account for both statistical and methodological heterogeneity as much as possible, and examining a smaller set of experiments that are similar in important ways should reduce levels of between-study heterogeneity.

I began my analysis with the assumption that the use of very different methodology produces estimates of very different effects. It follows from this assumption that one should not aggregate a large set of effect sizes derived from methodologically disparate experiments, but rather, that one should create smaller sets of effects sizes based on the similarity of the methods used. Therefore, I first identified the most frequently used manipulation and outcome tasks used in the sequential task paradigm. Second, I created a dataset consisting only of experiments in which both the outcome task and the manipulation task appear frequently in the literature. Third, I divided this dataset into smaller sets on the basis of the type of outcome tasks that were used. Finally, I

conducted meta-analytic tests of the depletion effect in each set, and applied methods for detecting and correcting for publication bias.

Data collection

The quality of a meta-analysis is completely dependent on the quality of the literature search that is used to locate the effect-sizes to be analyzed. For this reason, it is necessary to perform as thorough a search as possible of the extant literature (including peer-reviewed articles, dissertations and theses, conference presentations, and unpublished data). Including unpublished dissertations and theses, conference presentations, and experiments is utterly essential for producing trustworthy meta-analytic results. To that end, I conducted an exhaustive literature search using the following strategies: Searching of online databases (i.e., EBSCO, ISI Web of Science, and Proquest), and online lists of conference abstracts (i.e., for annual conferences for the Association for Psychological Science (APS) and the Society for Personality and Social Psychology (SPSP), personal communication with experts in the field, and issuing several calls for unpublished data through the listserv of SPSP. Additionally, all studies that were included in Hagger et al. are also included here. See appendix A for a description of the exact search process.

Inclusion criteria

For inclusion in the current analysis, an effect size must have been derived from an instance of the sequential task paradigm in which a measure of performance on the dependent task was available for participants in both the experimental and control conditions (i.e., from a true experiment). In some cases, tests of the depletion effect are purely correlational. For example, Vohs et al. (2008) found that the degree to which

shoppers reported making choices predicted persistence and performance on math problems. And in other cases, quasi-experimental methods are used to test the depletion effect. For example, Zelenski, Santoro, and Whelan (2012) assessed whether participants were chronically extraverted or chronically introverted and then required them to act either introverted or extraverted. Following this manipulation, participants completed the Stroop task. These studies were omitted.

Some experiments tested whether judgments, ratings, or responses to hypothetical situations or requests were affected by previous exertion of the active self. In one example, after performing a common depletion manipulation, participants were presented with a hypothetical situation in which they had been insulted and shoved by a stranger in a bar. Participants were then asked to rate how likely they would be to respond by assaulting the person with a beer bottle (DeWall, Baumeister, Stillman, & Gailliot, 2007). DeWall et al. (2007) examined whether the participant's rating of his or her willingness to resort to violence in this hypothetical situation was affected by the depletion manipulation. Because the participant is not faced with the reality of the situation, the validity of hypothetical measures is questionable (Baumeister, Vohs, & Funder, 2007); and therefore, tests in which the outcome variable is a judgment or a rating (such as the rating of the likelihood of a particular response to a hypothetical situation), should not be considered as comparable to tests of the depletion effect that are based in actual behavior.

As mentioned above, only effect sizes derived from instances of the sequential task paradigm in which both the manipulation task and the outcome task are frequently used tasks were included. Following Hagger et al., frequently used tasks will be defined

as those tasks that have been used in at least ten independent tests of the depletion effect (the ten instances must be either all as a manipulation or all as a dependent measure).

For studies in which an individual difference variable is thought to moderate the effect of depletion, only the main effect for the depletion manipulation was included. For studies in which an experimental factor is used as a moderator (e.g., administration of glucose to half of the participants; Gailliot, et al., 2007), I followed Hagger et al. (2010) in only including the effect size derived from the level of the moderator not thought to attenuate the depletion effect[1].

Coding

Effect sizes. Effect sizes took the form of bias-corrected standardized group mean differences (i.e., Hedge's *g*; Hedges, 1981)[2]. These were taken from all published and unpublished experiments that meet the above criteria. Hedge's *g* may be derived from any experiment that provides information about samples sizes, means, and sample standard deviations for the two groups. This may take the form of direct reporting of such statistics, but information about means and sample standard deviations can also be obtained from test statistics (i.e., *t* and *F* values) and from *p* values. When authors report only statistics for analyses that are elaborations on simple comparisons of means (i.e., paired-sample *t* tests, repeated measures analysis of variance, and analysis of covariance), additional information is needed to calculate *g*, such as the correlation between pre- and post-test scores or the correlation between the outcome and the covariate. When this information was available, *g* was calculated, and when it was not, the authors were contacted or, in some cases, an estimate was made (e.g., if a replication exists in which the necessary information is given, that information was applied to the experiment in

which it was missing). Formulas for computing *g* for these various cases are taken from Borenstein (2009).

In the case of incomplete information, two assumptions were made. First, if authors only reported the overall sample size, it was assumed that sample sizes were equal across groups (if the total sample size is odd, the remainder will be placed in the experimental group). Second, if authors only reported an effect as nonsignificant, or as "*NS*," it was assumed that the effect was zero.

If multiple effect size estimates were available from one outcome measure, a composite of each estimate was calculated. For example, there is no *a priori* reason to prefer reaction time to accuracy on the Stroop task as a measure of the depletion effect, and because both measures should reflect depletion, an aggregate of the two was computed using the method described by Gleisser and Olkin (2009). This method assumes that the two outcomes are correlated at the level of $r = 0.50$ level by default. When the true correlation between the multiple outcomes was not available, the default was used; however, if analogous experiments contained information about the correlation of interest, these values were used instead.

Variables of interest. Each effect size was coded for the following attributes, and in the case of significant between-study variation, these attribute codes were used as possible meta-analytic moderators. Four experiment-level variables were coded: publication status, source laboratory, the number of tasks used in the manipulation, and the number of tasks used as dependent measurements.

For publication status, experiments that published in peer-reviewed journals were coded as one. Experiments that were in press, under review, or being sent in for review were also coded as one. All other experiments were coded as zero.

For source laboratory, experiments were coded as one if one of the authors was associated with the Baumeister-Tice laboratory at Florida State University or a laboratory of a student from the Baumeister-Tice laboratory. Specifically, if any of the authors, or a committee member on a dissertation or master's, were Roy Baumeister, Diane Tice, Kathleen Vohs, Nathan DeWall, Mark Muraven, Brandon Schmeichel, or Matthew Gailliot, the experiment was coded as one. This variable was included because Hagger et al. included it in their original analysis, and because it may serve to account for significant variability in the depletion effect, either because of some phenomenon akin to allegiance bias (Leykin & DeRubeis, 2009) or because these researchers are particularly skilled at experimentally inducing the depletion effect.

For the number of tasks used in an experiment, if more than one manipulation or outcome task was used, the experiment was coded as a one. Otherwise, it was coded as a zero.

Statistical analysis

Inter-rater agreement. A trained graduate student and a trained undergraduate research assistant made all of the coding decisions regarding the number of manipulation tasks and the number of outcome tasks used. The same graduate student and I independently made each coding decision for the other variables discussed above. As recommended by Orwin and Vevea (2009), inter-rater agreement for nominal data, such as categorization of the number of manipulation tasks used, was calculated as Cohen's κ

coefficient (Cohen, 1960), and inter-rater agreement for continuous data was calculated using the Pearson correlation coefficient.

Random/mixed-effects model. In a RE model, when the τ^2 is not zero, the random effect represents variation from the mean of the true effects as a function of unobserved (or unmodeled) sources of heterogeneity, such as differences in methods used in individual experiments that is not coded and included in the model. Sources of heterogeneity can be further accounted for by including observed study characteristics in the model as moderators (i.e., τ^2 can be explained by modeling the association between study characteristics and the true effect sizes), resulting in the mixed effects (ME) linear model (so called because it includes both fixed and random effects). The ME model is given here as a single, combined model that integrates the two-level structure described above:

$$T_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \cdots + \beta_p X_{ip} + u_i + e_i ,$$

$$u_i + e_i \sim N(0, v_i^*),$$

and

$$v_i^* = \tau^2 + v_i ,$$

where β_0 is the intercept for the model, X_{i1}, …, X_{ip} are coded study characteristics for the i^{th} study, such as codes that represent the use of certain methods, and β_1, …, β_p are regression coefficients that represent the association between coded study characteristics and the true effect sizes. Together, u_i and e_i make up the total variance, v_i^*, in the observed effect size, T_i. That variance is partitioned into v_i, which is the observed variance of effect size i, and τ^2, which may now be thought of as the residual heterogeneity, or the amount of heterogeneity not accounted for by the model. Mixed-

effects models are fitted in a two-step approach, just as RE models (described above; Raudenbush, 2009).

RE models were run on each set of effect sizes derived using the same frequently used outcome task (for example, all studies that used some measure of accuracy or reaction time to Stroop trials as the dependent variable). When *Q* statistics were statistically significant, ME models including codes for the variables of interest were conducted (assuming that that these variables divided the sample up into subgroups containing more than one experiment—that is, if only one experiment in a sample was unpublished, then publication status was not used as a moderator). When predictors in a ME model were not statistically significant, and the overall *F*-test for moderators was not statistically significant, then the ME model was abandoned and it was concluded that none of the variables moderated the overall effect.

Publication Bias and Small Study Effects. As in Study One, the following methods were used to assess and correct for publication bias: The IC-index, the trim and fill, and the regression-based methods, PET and PEESE.

Chapter 5: Study Two Results

Data collection.

As described in the appendix, my literature search yielded 130 individual instances of the sequential task paradigm in which both the manipulation task and the outcome task were frequently occurring. Of the 130 experiments, 30 were excluded for analyses for one of three possible reasons: First, fourteen experiments did not contain enough information to code (all authors had been contacted about the missing information, but at the time that analyses were conducted, no reply had been received). Second, eight experiments included experiment-level moderators that did not have appropriate controls, and thus, no clear test of the depletion effect was available. And third, in eight additional experiments, the manipulation task was not used to manipulate use of self-control, but rather, as a means of inducing ego depletion in all the participants in the sample. Thus, the final sample was composed of exactly 100 independent instances of the sequential task paradigm (entries in the reference section preceded by an asterisk correspond to one of these 100 tests of the depletion effect).

Within the 100 experiments, there were ten categories of frequently occurring manipulation tasks (i.e., attention video, crossing letters, thought suppression, emotion video, Stroop, attention essay, food temptation, transcription, social exclusion, and working memory) and eight categories of frequently occurring outcome tasks (i.e., impossible anagrams, hand grip, impossible puzzles, food consumption, Stroop, possible anagrams, standardized tests, and working memory). These categories and the number of experiments in each category are listed and briefly described below.

Manipulation tasks

Attention video (*k* = 20). Participants watch a silent video during which stimuli occasionally appear. Participants in the control condition are given no instructions other than to watch the video, whereas participants in the experimental condition are told to ignore the stimuli when they appear. The video is usually of a woman being interviewed while words are displayed in the bottom right corner.

Crossing letters (*k* = 16). Participants are given sheets of paper with printed text. For the first page, participants are asked to cross out certain letters following certain rules. On the following page, participants in the experimental condition are given a different, more complex set of rules. Participants in the control condition continue on with the same rule.

Thought suppression (*k* = 13). Participants are asked to refrain from thinking about a certain topic. The most common version of this task is also known as "the white bear" paradigm because participants in the experimental condition are told that they can think about anything they want, except for a white bear. In contrast, participants in the control condition are told to think about whatever they want.

Emotion video (*k* = 14). Participants are shown an emotionally evocative video (e.g., a video of animals being harmed). Participants in the experimental condition are given instructions to regulate their emotions in some way (e.g., to either suppress or exaggerate them), whereas participants in the control condition are told to watch the video as they would any other video.

Stroop (*k* = 8). Participants are shown color words (e.g., the word yellow) printed in colored ink (e.g., blue) and told to name the color of the ink. Generally, participants in

the experimental condition are shown all incongruent trials (i.e., when the color of the ink does not match the color to which the word refers), whereas participants in the control condition are shown all congruent trials.

Attention essay (k = 11). Participants are asked to write about a topic (e.g., a recent vacation). Participants in the experimental condition are told that they cannot use some set of commonly occurring letters, usually *a* and *n*, while writing. In contrast, participants in the control condition are told that they cannot use uncommon letters, e.g., *q* and *z*.

Food temptation (k = 6).Participants in the experimental condition are told to resist the temptation to eat some type of food, usually a dessert. For example, participants are shown a plate of chocolates and a plate of radishes. Participants in the experimental condition are told to only eat radishes and keep from eating chocolates, whereas participants in the control condition are told to eat chocolate. Participants are commonly told that they are taking part in a taste test.

Transcription (k = 8). Participants are given a sheet of text and told to transcribe it. Participants in the experimental condition are told to transcribe the text without using certain keys, such as the space bar. Participants in the control condition are not given any additional instructions.

Social exclusion (k = 4). Participants are led to feel socially excluded. For example, while completing an ostensible task as a group, participants in the exclusion condition are told that no other participants wanted to work them. Participants in the control condition are typically included.

Working memory (*k* = 6). Participants in the experimental condition perform a task that is high in working memory load (e.g., remembering information while performing another task), whereas participants in the control condition perform a task that is relatively low in working memory load.

Outcome tasks

Food consumption (*k* = 14). The amount of food that participants consume is measured. For example, the number of ounces of ice cream consumed. Higher amounts of food are thought to be indicative of lower levels of self-control.

Hand grip (*k* = 13). Participants hold the arms of a hand grip closed for as long as possible (or hold a dynamometer at some percentage of their maximum grip strength). The length of time that participants are able to persist at this painful task is considered to indicate levels of self-control, with shorter times meaning lower levels of self-control.

Impossible anagrams (*k* = 18). Participants are given a set of anagrams to solve, some of which are designed to be impossible to solve. Persistence at this impossible task is thought to measure self-control, with less time spent (or lower numbers of attempts) being indicative of greater self-control.

Possible anagrams (*k* = 8). Participants are given a large set of anagrams and told to solve as many as possible. Lower numbers of solved anagrams are considered to be indicative of lower self-control.

Impossible puzzles (*k* = 14). Participants are asked to solve puzzles (e.g., tracing geometric shapes printed on paper without going back over the same line). Unbeknownst to participants, puzzles have been modified so that they are unsolvable. As with

impossible anagrams, persistence (either as time or as number of attempts) at this impossible task is used to index self-control.

Standardized tests (k = 13). Participants are given problems from some standardized test, typically the graduate record exam (GRE). The number or problems solve, the number of problems attempted, and the proportion of problems correct out of problems attempted are all used as an index of self-control (with worse performance being interpreted as lower self-control).

Stroop (k = 10). As described above, participants must identify the ink color of color words. Self-control is measured as the number of correct trials, as well as reaction time on trials (with slower reaction time meaning less self-control).

Working memory (k = 12). Participants perform some task designed to measure working memory. For example, the operation span task, in which participants must remember words or letters while solving simple math problems. Worse working memory performance (as indicated in a variety of ways, for example, fewer words recalled overall) is thought to indicate lower self-control.

Experiment-level characteristics

Tables 1 through 8 display the coded experiment-level characteristics for each experiment organized by the category of outcome task: the manipulation task used, the publication status of each experiment, the source laboratory for the experiment, whether multiple manipulation or multiple outcome tasks were used, the effect size derived from each experiment, its associated variance, and the sample sizes for the experimental and control groups are displayed. Note that, because two experiments included multiple outcome tasks that were both frequently used tasks, these two experiments are analyzed

in two samples: Experiment 4 from Muraven, Shmueli, and Burkley (2006) is included in both the impossible anagrams sample and the Stroop sample; Dewitte, Bruyneel, & Geyskens (2009) is included in both the food consumption sample and the possible anagrams sample.

Reliabilities for each coded characteristic are presented in Tables 1 through 8 as well. Due to a miscommunication between coders during the coding process, the codes for whether multiple manipulation tasks were present and whether multiple outcome tasks were present were accidentally combined. That is, a single experiment was only coded in terms of whether the raters agreed or disagreed on the type of task used for the manipulation and outcome tasks (i.e., each experiment was coded as a one when raters agreed on the coding of the tasks used and a zero when raters disagreed). In other words, separating out disagreement on presence of more than one manipulation task *or* more than one outcome task was not possible. Therefore, I only report the proportion of these experiments in which the raters agreed on the presence of multiple manipulation or outcome tasks (also called agreement rate [AR]).

This final set of experiments differed from the set of experiments analyzed by Hagger et al. in two notable respects. First, of the 100 experiments included in my sample, 33 were unpublished (in contrast to zero of the 198 in Hagger et al.'s sample). Second, more than half of the effect sizes in my sample were nonsignificant, in comparison to only 47 of the 198 in Hagger et al.'s dataset. These experiments are displayed in eight forest plots (organized by outcome task; figures 1 through 8) and eight contour-enhanced funnel plots (figures 9 through 16).

Primary analyses

Eight samples of experiments were created by dividing the 100 experiments into groups on the basis of outcome task category. The five statistical techniques described above (RE models, ME models, the IC-index, the trim and fill, and the PET/PEESE models) were applied to each sample.

Food consumption (k = 14). The estimate of the average overall depletion effect in the food consumption sample was g^* = 0.44 (95% confidence interval [-0.01, 0.89]), p = 0.06. This nonsignificant overall effect was qualified by an extreme degree of between-study heterogeneity, τ^2 = 0.52, Q = 96.75, p <0.001. The estimate for the amount of variance accounted for by factors other than sampling error was large, I^2 = 88.54% (95% confidence interval [79.1%, 91.0%]). See Table 9.

In this sample, it was possible to test two experiment-level characteristics—source laboratory and the number of outcome tasks—as possible meta-analytic moderators within a mixed-effects model. However, these predictors were not statistically significant. Furthermore, the test for residual heterogeneity was statistically significant after including the two predictors, Q_e = 87.68, p <0.001, suggesting that the average overall effect was not moderated by these factors (see Table 10).

The IC-index varied widely (from <0.01 to >0.99), which, given the width of the 95% confidence interval around the estimate of the overall average effect from the RE model, is not surprising. Therefore, the results from the IC-index were ambiguous, and it could not be definitively stated that too few nonsignificant findings were observed in this sample than are likely to exist. See Table 11.

No missing experiments were imputed when the trim and fill was applied, so the adjusted estimate was identical to that given by the RE model (Table 13). Not surprisingly then, b_1 for the PET model was nonsignificant, $b_1 = 2.38$, $p = 0.55$, indicating that there was a lack of funnel plot asymmetry in this sample. However, the overall estimate from the PET model was adjusted downward, $b_0 = -0.21$, $p = 0.83$ (see Table 12). The results from the PET model should be interpreted with caution since, in this particular sample, the average overall estimate from the random effects model was nonsignificant and accompanied by a very high level of between-study heterogeneity.

Based on these results, it appears that neither small study effects, such as publication bias, nor any of the coded experiment-level characteristics, moderate this nonsignificant average overall effect; however, based on the estimates of between-study heterogeneity, it is very likely that some other, unknown factor moderates the depletion effect in this sample.

Hand grip ($k = 13$). The estimate of the average overall depletion effect in the sample of experiments that used hand grip as an outcome was $g^* = 0.56$ (95% confidence interval [0.31, 0.81]), $p < 0.001$. The average overall effect was qualified by a relatively moderate degree of between-study heterogeneity, $\tau^2 = 0.09$, $Q = 26.85$, $p = 0.008$. The estimate for the amount of variance accounted for by factors other than sampling error was $I^2 = 55.73\%$ (95% confidence interval [16.7%, 76.0%]). See Table 9.

In the hand grip sample, because of a lack of variability in coded experiment-level characteristics (e.g., all but one experiment was published), it was not possible to use any of these characteristics as predictors in a mixed-effects model (see Table 2).

As in the food consumption sample, the IC-index varied widely (from 0.01 to 0.99). Therefore, the results from the IC-index were again ambiguous. See Table 11.

Four missing experiments were imputed when the trim and fill method was applied, and the adjusted estimate of the average overall effect was $g^* = 0.36$ (95% confidence interval [0.09, 0.63]), which represents a 36% reduction from the estimate provided by the RE model (see Table 13).

Results from the regression-based methods were similar to those of the trim and fill: b_1 for the PET model was significant, $b_1 = 4.76$, $p = 0.005$, which is indicative of funnel plot asymmetry, and the overall estimate from the PET model was adjusted downward and was nonsignificant, $b_0 = -0.76$, $p = 0.06$ (see Table 12). Based on the contour-enhanced funnel plot (Figure 10), it appears that the asymmetry is primarily in the area of nonsignificance, which is indicative of publication bias. Therefore, based on these results, it seems likely that the apparent evidence for the depletion effect in the hand grip sample is due to publication bias, and that, when statistically controlling for the influence of this bias, the average overall effect is no different from zero.

Impossible anagrams ($k = 18$). The estimate of the average overall depletion effect in the impossible anagrams sample was $g^* = 0.47$ (95% confidence interval [0.17, 0.77]), $p = 0.005$. The average overall effect was associated with a high degree of between-study heterogeneity, $\tau^2 = 0.25$, $Q = 67.4$, $p < 0.001$. The majority of the variance was accounted for by factors other than sampling error, $I^2 = 77.19\%$ (95% confidence interval [60%, 84.1%]). See Table 9.

For the impossible anagrams sample, it was possible to use each coded experiment-level characteristic as a predictor in a mixed-effects model. Source laboratory

and the presence of multiple manipulation tasks were the only significant predictors (see Table 10), so the other two predictors (publication status and the presence of multiple outcome tasks) were removed and the model was re-run: The intercept remained significant ($\beta = 0.44$, $p = 0.002$), as did both predictors, such that experiments associated with the Baumeister/Tice laboratory tended to be larger ($\beta = 0.75$, $p = 0.01$), and experiments in which more than one manipulation task was used tended to be smaller ($\beta = -0.71$, $p = 0.01$). However, despite these significant predictors, the test for residual heterogeneity was significant ($Q_e = 31.53$, $p = 0.001$), suggesting that the predictors did not account for all of the between-study heterogeneity (when moderators were included, τ^2 decreased from 0.25 to 0.07).

As in the previous two samples, the IC-index ranged widely (from 0.05 to 0.99). Thus, the results from the IC-index were again ambiguous, and it is not clear whether more nonsignificant findings than were observed in this sample are likely to exist. See Table 11.

Only one missing experiment was imputed when the trim and fill was applied, so the average overall effect was adjusted only slightly downward, $g^* = 0.41$ (95% confidence interval [0.13, 0.70]), $p = 0.005$ (Table 13), which represents a 13% decrease in magnitude from the estimate of the effect provided by the RE model. The test for funnel plot asymmetry from the PET model was nonsignificant, $b_1 = 0.26$, $p = 0.87$, and as in the food consumption sample, the overall estimate from the PET model was adjusted to be nonsignificant, despite a lack of funnel plot asymmetry, $b_0 = 0.39$, $p = 0.38$ (see Table 12). The non-significant b_0 for this sample seemed to be a distinctly different case from the non-significant b_0 for the food consumption sample, however. For the

impossible anagrams sample, the estimate of the overall effect from PET (b_0 = 0.39) was similar to the estimate from the RE model (g^* = 0.47). Thus, it would seem that the estimate of the overall effect from PET was nonsignificant, not because of a downward adjustment due to funnel plot asymmetry, but because of a widening of the confidence interval around b_0. In a case like this, it is not clear whether one should prefer the estimate of the overall effect from the RE model or from PET (see Moreno et al., 2012 for a similar example). Additionally, as in the food consumption sample, the results from the PET model should be interpreted with caution because of the high degree of between-study heterogeneity.

Based on these results, it is possible to conclude that there is evidence for the depletion effect in this sample. However, as with the food consumption sample, this overall average effect is apparently moderated by a variety of factors, such as the presence of more than one manipulation task (in this example, the sign of the depletion effect reverses, meaning that greater exertion of self-regulation leads to greater subsequent self-regulation performance, instead of lower subsequent performance as predicted by the limited strength model).

Possible anagrams (k = 8). The estimate of the average overall depletion effect in the sample of experiments that used possible anagrams as an outcome was g^* = 0.37 (95% confidence interval [0.11, 0.64]), p = 0.01. Unlike in the three previously described samples, there was relatively little between-study heterogeneity, τ^2 = 0.01, Q = 8.95, p = 0.26. The estimate for the amount of variance accounted for by factors other than sampling error was I^2 = 13.8% (95% confidence interval [0%, 63.8%]). See Table 9.

A mixed-effects model was not conducted in this sample because the *Q* statistic was not significant.

As in the previous three samples, the IC-index varied widely (from 0.06 to 0.99), and the results from the IC-index were ambiguous. See Table 11.

Three missing experiments were imputed when the trim and fill was applied. The estimate of the average overall effect when including these three imputed experiments was $g^* = 0.26$ (95% confidence interval [0.04, 0.48]), which represents a 30% reduction from the estimate provided by the RE model (see Table 13). Results from the regression-based methods were similar to those of the trim and fill: b_1 for the PET model was significant, $b_1 = 3.69$, $p = 0.017$, which is indicative of funnel plot asymmetry, and the overall estimate from the PET model was adjusted downward from that of the RE model and was nonsignificant, $b_0 = -0.62$, $p = 0.09$ (see Table 12).

The pattern of results for the possible anagrams sample was similar to the one observed in the hand grip sample: The apparent funnel plot asymmetry was primarily in the area of nonsignificance, which is indicative of publication bias (Figure 12). Therefore, as with the hand grip sample, it seems likely that any evidence for the depletion effect in the possible anagrams sample is due to publication bias, and that statistically controlling for this influence reduces the average overall effect to a value that does not differ from zero.

Impossible puzzles ($k = 14$). The estimate of the average overall depletion effect in the sample of experiments that used impossible puzzles as an outcome was $g^* = 0.79$ (95% confidence interval [0.53, 1.06]), $p < 0.001$. This was the largest estimate of the average overall depletion effect among the eight samples. This large effect was qualified

by a relatively moderate degree of between-study heterogeneity, $\tau^2 = 0.14$, $Q = 39.40$, $p < 0.001$. The estimate for the amount of variance accounted for by factors other than sampling error was $I^2 = 61.17\%$ (95% confidence interval [42.2%, 81.2%]). See Table 9.

In the impossible puzzle sample, it was possible to test the experiment-level characteristics of publication status and source laboratory as meta-analytic moderators by entering them as predictors in a mixed-effects model. However, both of these predictors were nonsignificant, whereas the test for residual heterogeneity was significant despite their inclusion, $Q_e = 34.25$, $p < 0.001$, suggesting that the average overall effect was not moderated by these factors (see Table 10).

As with the samples described above, the IC-index varied widely (from 0.06 to 0.99), and it was difficult to draw conclusions from these results. See Table 11.

Four missing experiments were imputed when the trim and fill was applied, and the adjusted estimate of the average overall effect was $g^* = 0.63$ (95% confidence interval [0.38, 0.88]), which represents a 20% reduction from the estimate provided by the RE model (see Table 13). Thus, the average overall depletion effect was only reduced somewhat, and the adjusted the effect was still of moderate magnitude. In contrast, results from the regression-based methods suggested that the overall effect was not different from zero. For the test of funnel plot asymmetry, b_1 for the PET model was significant, $b_1 = 3.01$, $p = 0.01$. The adjusted estimate of the average overall effect from the PET model was adjusted downward from that of the RE model and was nonsignificant, $b_0 = -0.16$, $p = 0.59$ (see Table 12). Based on the contour-enhanced funnel plot (Figure 13), it appears possible that the asymmetry is in the area of nonsignificance, which is indicative of publication bias; however, as can clearly be seen in Table 5, both published and

unpublished experiments tended to be non-zero and consistently of a medium to large magnitude. It is therefore somewhat difficult to claim that publication bias is the only possible cause of the observed asymmetry, and one should keep in mind the possibility of some other small-study effect. Regardless, once small-study effects have been statistically controlled (i.e., the estimate given in the PET model), the average overall effect is not different from zero.

Standardized tests ($k = 13$). The estimate of the average overall depletion effect in the sample of experiments that used standardized tests as an outcome was $g^* = 0.28$ (95% confidence interval [0.05, 0.52]), $p = 0.02$. Similar to the possible anagrams sample, there was relatively little between-study heterogeneity, $\tau^2 = 0.04$, $Q = 20.51$, $p = 0.06$. The estimate for the amount of variance accounted for by factors other than sampling error was $I^2 = 36.57\%$ (95% confidence interval [0%, 69.5%]). See Table 9.

As with the possible anagrams sample, a mixed-effects model was not conducted in this sample because the Q statistic was not significant.

As in the previous samples, the IC-index varied widely (from 0.04 to 0.85). However, despite this wide range, it is notable that even the maximum point of the IC-index does not provide strong evidence for the existence of other nonsignificant experiments outside of this sample. See Table 11. Thus, publication bias seems somewhat unlikely.

In concordance with the results for the IC-index, only one missing experiment was imputed when the trim and fill was applied, but it was imputed in the opposite direction as would have been expected if publication bias or other small-study effects were influencing the overall effect. The estimate of the average overall effect when

including this imputed experiment did not change from the estimate provided by the RE model, however (see Table 13). As with the food consumption and impossible anagrams samples, there was little evidence of funnel plot asymmetry: b_1 for the PET model was not significant, $b_1 = -0.17$, $p = 0.93$. Furthermore, and in line with the results from the trim and fill, the overall estimate from the PET model was actually adjusted slightly upward (though also brought to nonsignificance): $b_0 = 0.33$, $p = 0.52$ (see Table 12). This appeared to be a case similar to the PET results for the impossible anagram sample (see above).

The pattern of results in this sample can be thought of as evidence consistent with the depletion effect and inconsistent with the presence of publication bias (or other small-study effects). Importantly, the upward adjustment from the PET model is clearly due to the negative—rather than positive, as in the rest of the samples—correlation between standard error and effect size. This finding is most likely indicative of a lack of publication bias than it is of some underestimation of the depletion effect in this sample.

Stroop ($k = 10$). The estimate of the average overall depletion effect in the sample of experiments that use Stroop as an outcome task was $g^* = 0.38$ (95% confidence interval [0.12, 0.65]), $p = 0.01$. The average overall effect was associated with a moderate degree of between-study heterogeneity, $\tau^2 = 0.08$, $Q = 19.05$, $p = 0.02$. The majority of the variance was accounted for by factors other than sampling error, $I^2 = 53.54\%$ (95% confidence interval [3.2%, 76.9%]). See Table 9.

For the Stroop sample, it was possible to test publication status and source laboratory as predictors in a mixed-effects model. In this model, the intercept became nonsignificant ($\beta = 0.06$, $p = 0.69$) in the presence of publication status as a moderator

(which remained significant: $\beta = 0.49$, $p = 0.03$). The test for residual heterogeneity was not significant ($Q_e = 10.75$, $p = 0.22$), indicating that inclusion of publication status in the model accounted for all of the between-study heterogeneity. In other words, published results provide evidence of the depletion effect, whereas unpublished results do not—a result that was highly suggestive of publication bias.

Four missing experiments were imputed when the trim and fill was applied. The estimate of the average overall effect when including these four imputed experiments was $g^* = 0.17$ (95% confidence interval [-0.09, 0.43]), which represents a 52% reduction from the estimate provided by the RE model (see Table 13). Importantly, the adjusted estimate was not statistically significant, meaning that after accounting for possible publication bias via the trim and fill, the estimate of the overall average effect was not appreciably different from zero. Results from the regression-based methods mirrored those of the trim and fill: b_1 for the PET model was significant, $b_1 = 4.21$, $p = 0.012$, which is indicative of funnel plot asymmetry, and the overall estimate from the PET model was adjusted downward from that of the RE model and was nonsignificant, $b_0 = -0.72$, $p = 0.06$ (see Table 12).

The pattern of results for this sample, as with the results from the hand grip and possible anagrams samples, are indicative of publication bias: The apparent funnel plot asymmetry was primarily in the area of nonsignificance (Figure 15), and results from the mixed-effects model clearly suggest a discrepancy in results that was attributable to publication status. Therefore, as with the hand grip and possible anagram samples, it seems likely that any evidence for the depletion effect in the Stroop sample is due to

publication bias, and that statistically controlling for this influence reduces the average overall effect to a value that does not differ from zero.

Working memory (k = 12). The estimate of the average overall depletion effect in the sample of experiments that used working memory as an outcome was g^* = 0.33 (95% confidence interval [0.11, 0.56]), p = 0.008. The effect was associated with a relatively moderate degree of between-study heterogeneity, τ^2 = 0.05, Q = 21.9, p = 0.03. The estimate for the amount of variance accounted for by factors other than sampling error was I^2 = 47.83% (95% confidence interval [2.5%, 74.18%]). See Table 9.

In the working memory sample, it was possible to test the experiment-level characteristics of source laboratory and the presence of multiple outcome tasks as meta-analytic moderators by entering them as predictors in a mixed-effects model. However, both of these predictors were nonsignificant, whereas the test for residual heterogeneity was significant despite their inclusion, Q_e = 16.63, p = 0.05, suggesting that the average overall effect was not moderated by these factors (see Table 10).

As with the samples described above, the IC-index varied widely (from 0.21 to 0.99). Therefore, the results from the IC-index were again ambiguous. See Table 11.

No missing experiments were imputed when the trim and fill was applied, so the overall estimate was not adjusted by this method (see Table 13). However, results from the regression-based methods were indicative of the influence of small-study effects, b_1 = 3.87, p = 0.03, and the average overall effect was adjusted downward to be nonsignificant, b_0 = -0.61, p = 0.12 (see Table 12). Based on the contour-enhanced funnel plot (Figure 16), the asymmetry is not clearly in the area of nonsignificance. It therefore seems likely that publication bias is not the only possible cause of the observed

asymmetry, and instead, some other small-study effect may be creating an association between effect size and standard error. Regardless, once small-study effects have been statistically controlled (i.e., the estimate given in the PET model), the average overall effect is not different from zero.

Summary. In all but one sample (i.e., food consumption), the estimate of the average overall effect from the RE model was statistically significant. However, in all but two samples (i.e., possible anagrams and standardized tests), the average overall effect was qualified by moderate to extreme between-study heterogeneity, suggesting that a single summary estimate does not best represent the depletion effect. For the impossible anagrams sample, the between-study heterogeneity was partially explained by two experiment-level characteristics: (a) whether experiments were conducted by authors associated with the Baumeister/Tice laboratory (such that the estimate of the average overall effect for these experiments increased by 0.79 standard deviation units), and (b) whether experiments included more than a single manipulation task (such that the estimate of the average overall effect for these experiments decreased by -0.72 standard deviation units). For the Stroop sample, all of the between-study heterogeneity was explained by the publication status of the experiments, and the estimate of the average overall effect for unpublished studies—that is, the intercept in the mixed-effect model—was 0.06.

In five of the samples (hand grip, possible anagrams, impossible anagrams, Stroop, and working memory), between-study heterogeneity appeared to be at least partially due to the presence of small-study effects. In three of these five samples (i.e., hand grip, Stroop, and possible anagrams), the small-study effect in question appeared to

be publication bias, whereas the exact mechanism was less obvious in the other two samples. Regardless, in each of the five samples where evidence for an influence of small study effects was found, controlling for this influence reduced the estimate of the overall average effect to nonsignificance.

Finally, findings from the trim and fill method and from the IC-index were ambiguous. Both methods offer some evidence for the possibility of publication bias, but the large degree of between-study heterogeneity in the samples, as well as the wide confidence intervals around the estimate of the average overall effect from the RE models, make results from the trim and fill and IC-index difficult to interpret.

Table 14 has been provided as a summary of how one may interpret the evidence for the depletion effect as proposed in the limited strength model in the face of the results described above. This table includes four key questions that can be applied to each sample: (1) Is the average overall depletion effect statistically significant?; (2) After imputing experiments that are potentially missing due to publication bias, is the overall average depletion effect still significant?; (3) Is the overall average depletion effect still significant after correcting for small-study effects?; (4) When the overall average depletion effect is moderated by an observed experiment-level characteristics, is the effect significant at all levels of the moderator(s)? If, for a given sample, the answer to each of these questions is yes, then it is possible to argue that my results represent evidence for the depletion effect in that sample. As can be seen from Table 14, however, it is only possible to interpret results from the standardized tests sample as evidence in favor of the depletion effect.

Chapter 4: Discussion

As mentioned above, Hagger et al. described the results of their meta-analysis as "…demonstrating that the ego-depletion effect exists, its associated confidence intervals do not include trivial values, and it is generalizable across spheres of self-control" (p. 515). My results (both from Study One and Study Two) contradict each aspect of this statement. Indeed, the most certain conclusion that can be drawn from my results is that the depletion effect is *not* consistently observable across a variety of experimental contexts.

In Study Two, the typically high degree of statistical heterogeneity in the majority of the samples implies that the depletion effect varies widely in magnitude, and in most cases, is moderated by factors such as small-study effects (e.g., publication bias) or some experiment-level characteristic. In the second most heterogeneous sample, impossible anagrams, some of the between-study heterogeneity appears to be due to an increase in the depletion effect as a function some of the experiments being conducted in association with the Baumeister/Tice laboratory, although this was the only statistically significant instance of an apparent allegiance effect across the eight meta-analytic subsamples, so it would be ill-advised to conclude that this finding characterizes the literature on the limited strength model in general. Indeed, even this single result may be spurious.

In the same sample, additional heterogeneity is due to the use of multiple manipulation tasks (experiments 2 and 3 from Holmqvist [2008] and experiment 2 from Converse and Deshon, 2009). Interestingly, of the 100 experiments in my meta-analytic sample, four included multiple manipulation tasks (experiments 2 and 3 from Converse and Deshon [2009]; Carter and McCullough [2013], experiments 2 and 3 from Holmqvist

[2008]), and in each of these experiments, the estimated effect size was less than zero. Thus, performing multiple manipulation tasks may lead to higher subsequent efforts at self-regulation, rather than lower subsequent efforts, as one would predict based on the limited strength model. This is consistent with work by Converse and DeShon (2009), who, drawing on the literature of Learned Industriousness (Eisenberger, 1992), proposed that increasing the intensity of the initial self-regulatory effort by increasing the number of tasks participants are required to perform results in improved subsequent self-regulatory effort (i.e., the opposite of the depletion effect). Such a non-linear relationship between previous self-regulation and subsequent self-regulation is in stark contrast to the linear relationship that is at the core of the limited strength model (e.g., Vohs, Baumeister, & Schmeichel, 2012).

In at least three samples (hand grip, possible anagrams, and Stroop), the evidence for the depletion effect seems very likely due to publication bias. There is clear evidence of funnel plot asymmetry in each case, and a visual inspection of the contour-enhanced funnel plots in Figures 9, 12, and 15 suggests that the missing experiments are likely to be from the nonsignificant region of the funnel. Furthermore, of the 31 experiments that make up these three samples, only four were unpublished (one in the hand grip sample, Neale-Lorello [2009], and three in the Stroop sample, Cesario [2011], Myers [2010], and Pond, DeWall, Carter, & McCullough [2013]). Each of these unpublished experiments reported nonsignificant results, and each were among the most precise experiments in their respective samples (Neale-Lorello [2009] was the fourth most precise in the handgrip sample; Cesario [2011]; Myers [2010]; and Pond, et al. [2011] were the first, second, and fourth most precise estimates in the Stroop sample, respectively). It was only

possible to test publication status as a moderator in a mixed-effects model for the Stroop sample, and, as described above, the results of this test showed that controlling for publication status brought the overall effect down to approximately zero ($b = 0.06$). These results were mirrored by those from the PET and PEESE models, which in the case of all three samples, produced adjusted average overall effects that were no different from zero.

In respect to the possibility of publication bias, it is worth noting that the IC-index did not uniformly indicate that the number of significant findings observed was "too incredible" given the average power of the samples. One might argue that this constitutes evidence against the possibility of publication bias; however, the results from the IC-index varied widely, and in the cases in which the effect size estimate used to calculate power was taken from the lower limit of the confidence interval from the RE model, the IC-index was unanimously high (IC-index of 0.85 in the standardized tests sample and 0.99 in all other samples). It is difficult to draw any conclusions from the results of the IC-index, whether in favor of the presence of publication bias or not.

In the impossible puzzles and working memory samples, the apparent funnel plot asymmetry is not clearly due to publication bias. In the impossible puzzles sample, effect size estimates were generally positive and large, regardless of publication status, so publication bias is not an obvious explanation for the observed asymmetry (though, based on the contour-enhanced funnel plot [Figure 13], one should not rule out publication bias). For the working memory sample, the majority of the experiments fell within the nonsignificant range, so selection for significant findings seems unlikely. For these two samples, instead of publication bias, a better explanation might be some other form of

small study effect. Of course, given my efforts to ensure that methodological heterogeneity was minimized in the samples, and given the relatively straightforward nature of the sequential task paradigm, it is hard to imagine what form such a small study effect would take.

Examining what factors tend to covary with sample size for the sequential task paradigm, and thus, potentially result in a larger depletion effect, may be an avenue for future work. However, these particular results are consistent with the overall theme of my findings—in seven of the eight samples examined, evidence for the depletion effect must be qualified by the apparent moderation of the depletion effect by other factors. The one sample that was an exception, standardized tests, therefore, might represent an instance in which the depletion effect is consistently different from zero (though it is important to note that the lower limit of the 95% confidence interval on this estimate was $g = 0.05$, which is a very small effect). The fact that this pattern of results only appeared in one of eight samples is a direct contradiction of the limited strength model, which holds that any task that requires use of the active self should evince the depletion effect. Therefore, the depletion effect may exist in some specialized form, but the present results strongly suggest that the general form of the depletion effect proposed in the limited strength model does not exist.

Conclusion. For this dissertation, I sought to produce an estimate of the depletion effect in which even skeptical readers could feel confident. However, based on my findings, such an estimate does not appear to exist. The evidence for the depletion effect was either very likely due to publication bias, or difficult to interpret as a single

underlying effect size due to the presence of moderators of the effect (e.g., unidentified small-study effects or the number of manipulation tasks used).

Future research may be conducted to elucidate the exact instances in which the depletion effect does or does not occur; but, given the dangers of publication bias, it is critical that these future experiments are pre-registered in some way and that every effort be made to make these results accessible. Additionally, given the apparent existence of small-study effects that were not clearly related to publication bias, it is of particular importance to collect very large samples when studying the depletion effect. If such pre-registered, large studies are indeed conducted in the future, it would be simple to add these experiments to the samples I have reported on here, thereby helping to clarify the true nature of the depletion effect.

One should note the distinction between phenomenon and experimental protocol—that is, it is possible that previous acts of self-control do negatively affect subsequent acts of self-control, but that this phenomenon cannot be measured by the sequential task paradigm. However, at the time of this writing, and based on the available body of evidence that has been acquired through use of the sequential task paradigm and reviewed here, claiming that the limited strength model is a useful explanation for the failure of self-control seems unwarranted.

References

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.* Washington, DC: American Psychiatric Association.

*Barber, S. J., & Rajaram, S. (2011). Collaborative memory and part-set cueing impairments: The role of executive depletion in modulating retrieval disruption. *Memory, 19*(4), 378-397.

Bauer, I. M. & Baumeister, R. F. (2011). Self-regulatory strength. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 64–82). New York, NY: Guilford Press.

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

*Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Twenge, J. M. (2005). Social exclusion impairs self-regulation. *Journal of Personality and Social Psychology, 88*(4), 589-604. doi: 10.1037/0022-3514.88.4.589

Baumeister, R. F., Muraven, M., & Tice, D. M. (2000). Ego depletion: A resource model of volition, self-regulation, and controlled processing. *Social Cognition,* 18, 130–150.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science,* 2, 396–403.

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

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 279–293). New York, NY: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and RE models for meta-analysis. *Research Synthesis Methods,* 1, 97–111.

*Boucher, H. C., & Kofos, M. N. (2012). The idea of money counteracts ego depletion effects. *Journal of Experimental Social Psychology, 48*(4), 804-810.

*Bray, S., Martin, G. K., & Woodgate, J. (2011). Self-regulatory strength depletion and muscle-endurance performance: a test of the limited-strength model in older adults. *Journal of Aging and Physical Activity, 19*(3), 177-188.

*Bray, S. R., Martin Ginis, K. A., Hicks, A. L., & Woodgate, J. (2008). Effects of self-regulatory strength depletion on muscular performance and EMG activation. *Psychophysiology, 45*(2), 337-343.

Burton, V. S., Cullen, F. T., Evans, T. D., Alarid, L. F., & Dunaway, R. G. (1998). Gender, self-control, and crime. *Journal of Research in Crime & Delinquency,* 35, 123–147.

*Carter, E. C., & McCullough, M. E. (2013). After a Pair of Self-Control-Intensive Tasks, Sucrose Swishing Improves Subsequent Working Memory Performance. Manuscript sent in for publication.

Carver, C. S. & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical and health psychology. *Psychological Bulletin*, 92, 111–135.

Carver, C. S. & Scheier, M. F. (1998). *On the self-regulation of behavior.* New York, NY: Cambridge University Press.

Carver, C. S. & Scheier, M. F. (2011). Self-regulation of action and affect. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 3–21). New York, NY: Guilford Press.

*Cesario, J. (2011). Glucose drinks and self-control (stroop). Retrieved from http://psychfiledrawer.org/replication.php?attempt=MTIw.

Chan, A., Hrobjartsson, A., Haahr, M. T., Peter, C. G., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials. Comparison of protocols to published articles. *Journal of the American Medical Association*, 291, 2457–2465.

*Christiansen, P., Cole, J. C., & Field, M. (2012). Ego depletion increases ad-lib alcohol consumption: investigating cognitive mediators and moderators. *Experimental and Clinical Psychopharmacology, 20*(2), 118-128.

*Clarkson, J. J., Hirt, E. R., Jia, L., & Alexander, M. B. (2010). When perception is more than reality: the effects of perceived versus actual resource depletion on self-regulatory behavior. *Journal of Personality and Social Psychology, 98*(1), 29-46.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

*Converse, P. D., & Deshon, R. P. (2009). A tale of two tasks: reversing the self-regulatory resource depletion effect. *The Journal of Applied Psychology, 94*(5), 1318-1324.

Cooper, H., Hedges, L.V., & Valentine, J.C. (2009). *The handbook of research synthesis and metaanalysis* (2nd edition). New York, NY: Russell Sage Foundation.

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, 62-76.

DeWall, C. N., Baumeister, R. F., Gailliot, M. T., & Maner, J. K. (2008). Depletion makes the heart grow less helpful: Helping as a function of self-regulatory energy and genetic relatedness. *Personality and Social Psychology Bulletin,* 34, 1653–1662.

*DeWall, C. N., Baumeister, R. F., & Vohs, K. D. (2008). Satiated with belongingness? Effects of acceptance, rejection, and task framing on self-regulatory performance. *Journal of Personality and Social Psychology, 95*(6), 1367-1382.

*Dewitte, S., Bruyneel, S., & Geyskens, K. (2009). Self-regulating enhances self-regulation in subsequent consumer decisions involving similar response conflicts. *Journal of Consumer Research, 36*(3), 394-405.

*Dingemans, A. E., Martijn, C., Jansen, A. T., & van Furth, E. F. (2009). The effect of suppressing negative emotions on eating behavior in binge eating disorder. *Appetite, 52*(1), 51-57.

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

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

*Dvorak, R. D., & Simons, J. S. (2009). Moderation of resource depletion in the self-control strength model: differing effects of two modes of self-control. *Personality and Social Psychology Bulletin, 35*(5), 572-583.

*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., Davey Smith, G., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal,* 315, 629–634.

Eisenberger, R. (1992). Learned industriousness. *Psychological Review, 99*(2), 248–267.

Feldman, S. C., Martinez-Pons, M., & Shaham, D. (1995). The relationship of self-efficacy, self-regulation, and collaborative verbal behavior with grades: Preliminary findings. *Psychological Reports,* 77, 971–978.

*Friese, M., Hofmann, W., & Wanke, M. (2008). When impulses take over: moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *The British Journal of Social Psychology, 47*(3), 397-419.

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, 49–62.

*Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., Brewer, L. E., & Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology,* 92, 325–336.

*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.

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

Gleser & Olkin (2009). Stochastically dependent effect sizes. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 357–377). New York, NY: Russell Sage Foundation.

*Gohar, D. (2011). *Improving Self-Regulation: Increasing Resistance to Ego Depletion and Counteracting its Negative Effects* (Unpublished master's thesis). University of Pennsylvania, Philadelphia, PA.

Gottfredson, M. R. & Hirschi, T. (1990). *A general theory of crime.* Stanford, CA: Stanford University Press.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, 136, 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.

*Hedgcock, W. M., Vohs, K. D., & Rao, A. R. (2012). Reducing self-control depletion effects through enhanced sensitivity to implementation: Evidence from fMRI and behavioral studies. *Journal of Consumer Psychology, 22*(4), 486-495.

Hedges, L. V. (1981). "Distribution theory for Glass's estimator of effect size and related estimators". *Journal of Educational Statistics,* 6, 107–128.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 37–47). New York, NY: Russell Sage Foundation.

Higgins, J. P. T. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539–58.

*Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. *Journal of Experimental Social Psychology, 43*(3), 497-504.

*Holmqvist, M. E. (2008). *The influence of state and trait energy on self-regulatory behavior* (Unpublished doctoral dissertation). University of Saskatchewan, Saskatoon, Saskatchewan.

*Imhoff, R., Schmidt, A. F., Dislich, F. (2011). *Is there a downside to self-control? Ironic effects of trait self-control after ego depletion.* Manuscript submitted for publication.

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

Ioannidis, J. P. A. & Trikalinos, T. A. (2007b). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal,* 8, 1091–1096.

Ioannidis, J. P. A. & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials,* 4, 245–253.

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, 1041−1045.

*Klaphake, S. L. (2011). *Depletion and Replenishment: Exploring Self-Regulation Resource Depletion, Activities that Replensih the Resource, and the Corresponding Effects on Mood* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Knapp, G. & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine,* 22, 2693–2710.

Latham, L. L., & Perlow, R. (1996). The relationship of client-directed aggressive and nonclient-directed aggressive work behavior with self-control. *Journal of Applied Social Psychology,* 26, 1027–1041.

*Lattimore, P., & Maxwell, L. (2004). Cognitive load, stress, and disinhibited eating. *Eating Behaviors, 5*(4), 315–324.

Leykin, Y. & DeRubeis, R. J. (2009). Allegiance in psychotherapy outcome research: Separating association from bias. *Clinical Psychology: Science and Practice,* 16, 54–65.

Light, R. J., & Pillemer, D. B. (1984). Summing up: The science of reviewing research. Cambridge, MA: Harvard University Press.

*Litvin, E. B., Kovacs, M. A., Hayes, P. L., & Brandon, T. H. (2012). Responding to tobacco craving: experimental test of acceptance versus suppression. *Psychology of Addictive Behaviors, 26*(4), 830-837.

Madden, G. J. & Bickel, W. K. (2010). *Impulsivity: The behavioral and neurological science of discounting*. Washington, DC: American Psychological Association.

Mann, J. J., Waternaux, C., Gretcher, L. H., & Malone, K. M. (1999). Toward a clinical model of suicidal behavior in psychiatric patients. *American Journal of Psychiatry,* 156, 181–189.

*Martijn, C., Tenbült, P., Merckelbach, H., Dreezens, E., & de Vries, N. K. (2002). Getting a grip on ourselves: Challenging expectancies about loss of energy after self-control. *Social Cognition, 20*(6), 441-460.

McGuire, J., & Broomfield, D. (1994). Violent offenses and capacity for self-control. *Psychology Crime & Law,* 2, 117–123.

Mischel, W., Shoda, Y., & Peake, P. K. (1988). The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality and Social Psychology,* 54, 687–696.

*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.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, 9, 1–17.

Moreno, S. G., Sutton, A. J., Thompson, J. R., Ades, A. E., Abrams, K. R., & Cooper, N. J. (2012). A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity. *Statistics in Medicine*, 31, 1407–1417.

Moeller, F. G., Barratt, E. S., Dougherty, D. M., Schmitz, J. M., & Swann, A. C. (2001) Psychiatric aspects of impulsivity. *American Journal of Psychiatry,*158, 1783–1793.

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

*Muraven, M., Collins, R. L., & Nienhaus, K. (2002). Self-control and alcohol restraint: An initial application of the self-control strength model. *Psychology of Addictive Behaviors, 16*(2), 113-120.

*Muraven, M., Shmueli, D., & Burkley, E. (2006). Conserving self-control strength. *Journal of Personality and Social Psychology, 91*(3), 524-537.

*Muraven, M., & Slessareva, E. (2003). Mechanisms of self-control failure: Motivation and limited resources. *Personality and Social Psychology Bulletin, 29*(7), 894-906.

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

*Murtagh, A. M., & Todd, S. A. (2004). Self-regulation: A challenge to the strength model. *Journal of Articles in Support of the Null Hypothesis, 3*(1), 19-51.

*Myers, J. (2010). *Self-control: Is glucose a constraint or an input?* (Unpublished master's thesis). University of Pennsylvania, Philadelphia, PA.

*Neale-Lorello, D. G. (2009). *Mindfulness as a Self-Regulatory Act: Exploring the Relationship of Mindfulness to Ego Depletion* (Unpublished master's thesis). American University, Washington, DC.

Nigg, J. T., Quamma, J. P., Greenberg, M. T., & Kusche, C. A. (1999). A two-year longitudinal study of neuropsychological and cognitive performance in relation to behavioral problems and competencies in elementary school children. *Journal of Abnormal Child Psychology*, 27, 51–63.

*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.

Orwin, R. G. & Vevea, J. L. (2009). In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 177–203). New York, NY: Russell Sage Foundation.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, 61, 991–996.

Pond, R., DeWall, C. N., Carter, E. C., & McCullough, M. E. (2013). Religious priming and ego depletion. Unpublished data.

Powers, W. T. (1973). Feedback: Beyond behaviorism. *Science*, 179, 351–356.

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W. (2009). Analyzing effect sizes: RE models. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). New York, NY: Russell Sage Foundation.

Romal, J. B., & Kaplan, B. J. (1995). Difference in self-control among spenders and savers. *Psychology—A Quarterly Journal of Human Behavior,* 32, 8–17.

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, 59, 464–468.

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication bias and meta-analysis: Prevention, assessments and adjustments.* Chichester: John Wiley & Sons.

*Ruci, L. (2010). *Pro-social personality traits and helping motivations: Using the concept of ego depletion in distinguishing between instrinsically and extrinsically motivated helping* (Unpublished doctoral dissertation). Carleton University, Ottawa, Canada.

Rücker, G., Carpenter, J. R., & Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. Biometrical Journal, 52, 351–368.

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology,* 8, 1–9.

*Sato, T., Harman, B. A., Donohoe, W. M., Weaver, A., & Hall, W. A. (2010). Individual differences in ego depletion: The role of sociotropy-autonomy. *Motivation and Emotion, 34*(2), 205-213.

Scherschel, H. (2011). *Does religiousness protect against depletion effects?* Unpublished manuscript.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple study articles. *Psychological Methods,* 17, 551–566.

*Schmeichel, B. J. (2005). *Ego depletion, working memory, and the executive function of the self* (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

*Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *Journal 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.

*Seeley, E. A., & Gardner, W. L. (2003). The "selfless" and self-regulation: The role of chronic other-orientation in averting self-regulatory depletion. *Self and Identity*, *2*(2), 103-117.

*Segerstrom, S. C., & Nes, L. S. (2007). Heart rate variability reflects self-regulatory strength, effort, and fatigue. *Psychological Science, 18*(3), 275-281.

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology,* 26, 978–986.

*Smith, R. W. (2002). *Effects of relaxation on self-regulatory depletion* (Unpublished doctoral dissertation). Case Western Reserve University, Cleveland, OH.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. Oxford Bulletin of Economics and Statistics, 70, 103-127.

Stanely, T. D. & Doucouliagos, H. (2011). Meta-regression approximations to reduce publication selection bias. *Economics Series,*4.

Stern, J. M. & Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal,* 315, 640–645.

*Stillman, T. F., Tice, D. M., Fincham, F. D., & Lambert, N. M. (2009). The psychological presence of family improves self-control. *Journal of Social and Clinical Psychology, 28*(4), 498-529.

Sutton, A. J. (2009). Publication Bias. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 435–452). New York, NY: Russell Sage Foundation.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72, 271–322.

Terrin, N., Schmid, C. H., Lau, J., Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine,* 22, 2113–26.

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology,* 58, 894–901.

*Tyler, J. M., & Burns, K. C. (2009). Triggering conservation of the self's regulatory resources. *Basic and Applied Social Psychology, 31*(3), 255-266.

*Uziel, L., & Baumeister, R. F. (2012). The effect of public social context on self-control: depletion for neuroticism and restoration for impression management. *Personality and Social Psychology Bulletin, 38*(3), 384-396.

*vanDellen, M. R., Hoyle, R. H., & Miller, R. (2012). The regulatory easy street: Self-regulation below the self-control threshold does not consume regulatory resources. *Personality and Individual Differences, 52*(8), 898-902.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the RE model. *Journal of Educational and Behavioral Statistics,* 30, 261–293.

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

*Vohs, K. D., Baumeister, R. F., Mead, N. L., Ramanathan, S., Schmeichel, B. J. (2013). *Engaging in self-control heightens urges and feelings.* Unpublished manuscript.

Vohs, K. D., Baumeister, R. F., Nelson, N. M., Rawn, C. D., Twenge, J. M., Schmeichel, B. J., & Tice, D. M. (2007). Making choices impairs subsequent self-control: A limited resource account of decision making, self-regulation, and active initiative. *Personality Processes and Individual Differences,* 94, 883–898.

Vohs, K.D., Baumeister, R.F., Schmeichel, B.J. (2012). Motivation, personal beliefs, and limited resources all contribute to self-control. *Journal of Experimental Social Psychology,* 48, 943–947.

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

Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *American Psychological Society, 2*(3).

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

*Wan, W. (2007). *A monitoring model for understanding the regulatory depletion effect* (Unpublished doctoral dissertation). Northwestern University, Evanston, IL.

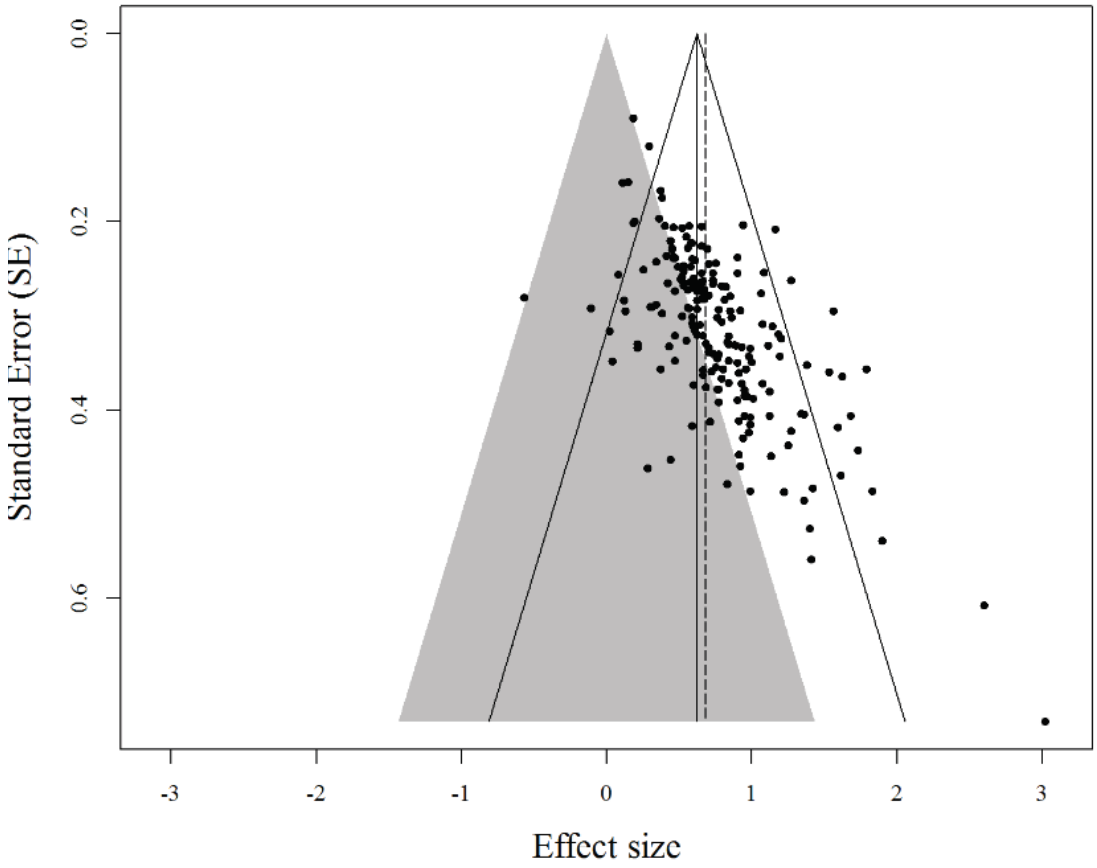
Wiener, N. (1948). *Cybernetics: Control and communication in the animal and the machine.* Cambridge, MA: M.I.T. Press.

Wolfe, R. N. & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational & Psychological Measurement,* 55, 177–185.

*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.

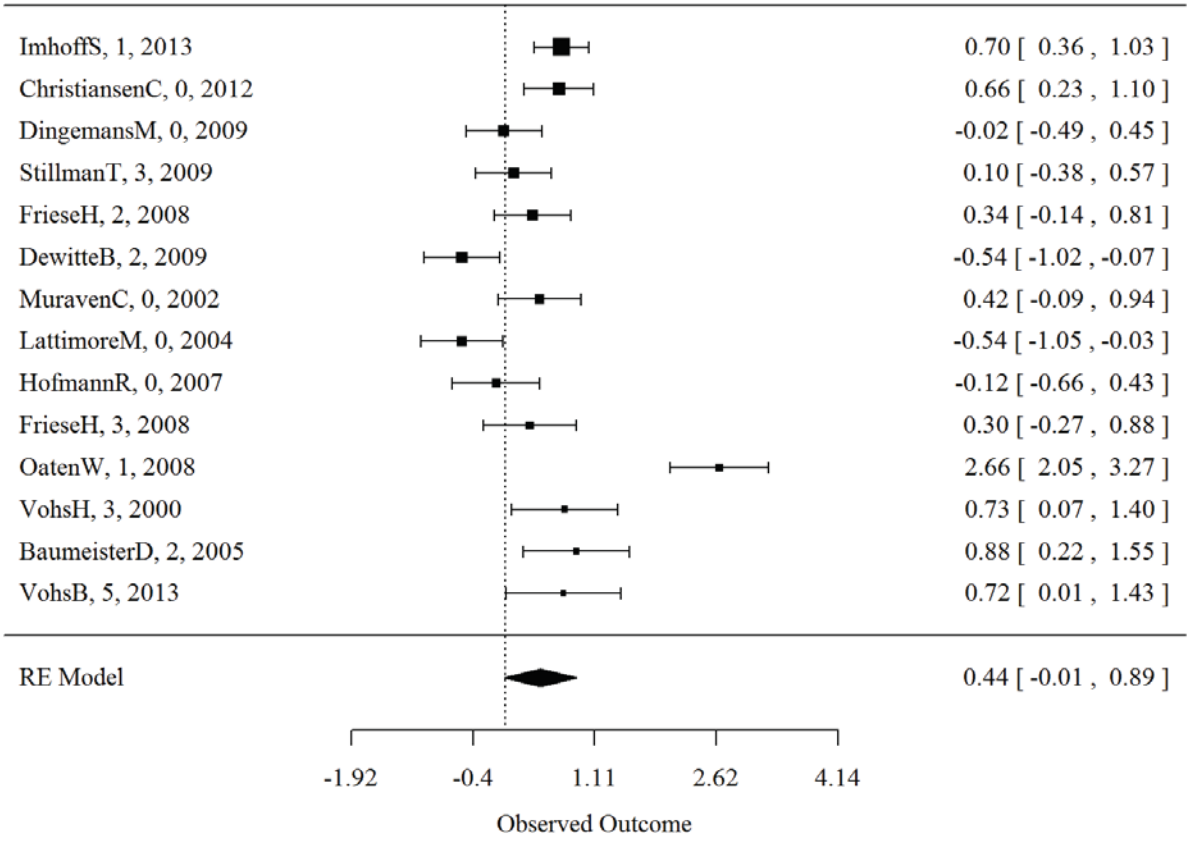
Zelenski, J. M., Santoro, M. S., & Whelan, D. C. (2012). Would introverts be better off if they acted more like extraverts? Exploring emotional and cognitive consequences of counterdispositional behavior. *Emotion,* 12, 290-303.

Figure 1. Contour-enhanced funnel plot of the experiments in Hagger et al.'s sample.



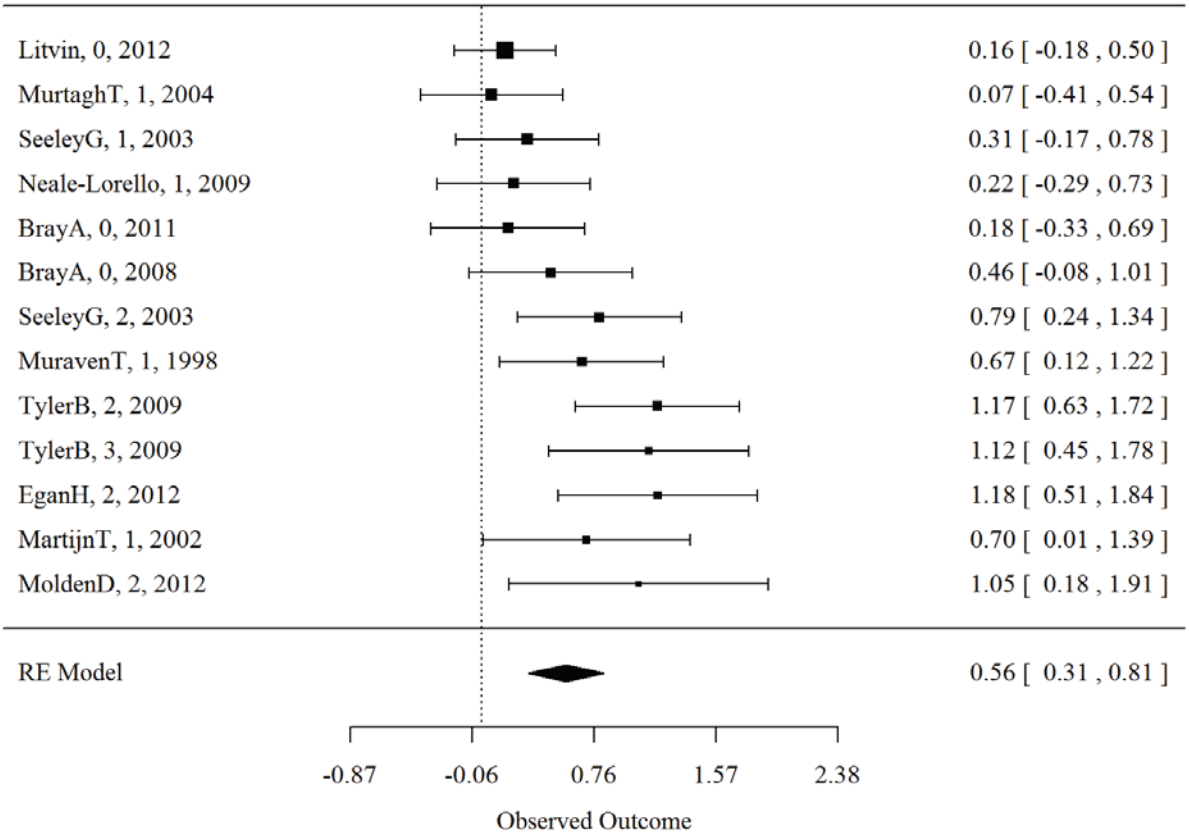
Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 2. Forest plot of the experiments in the food consumption sample.



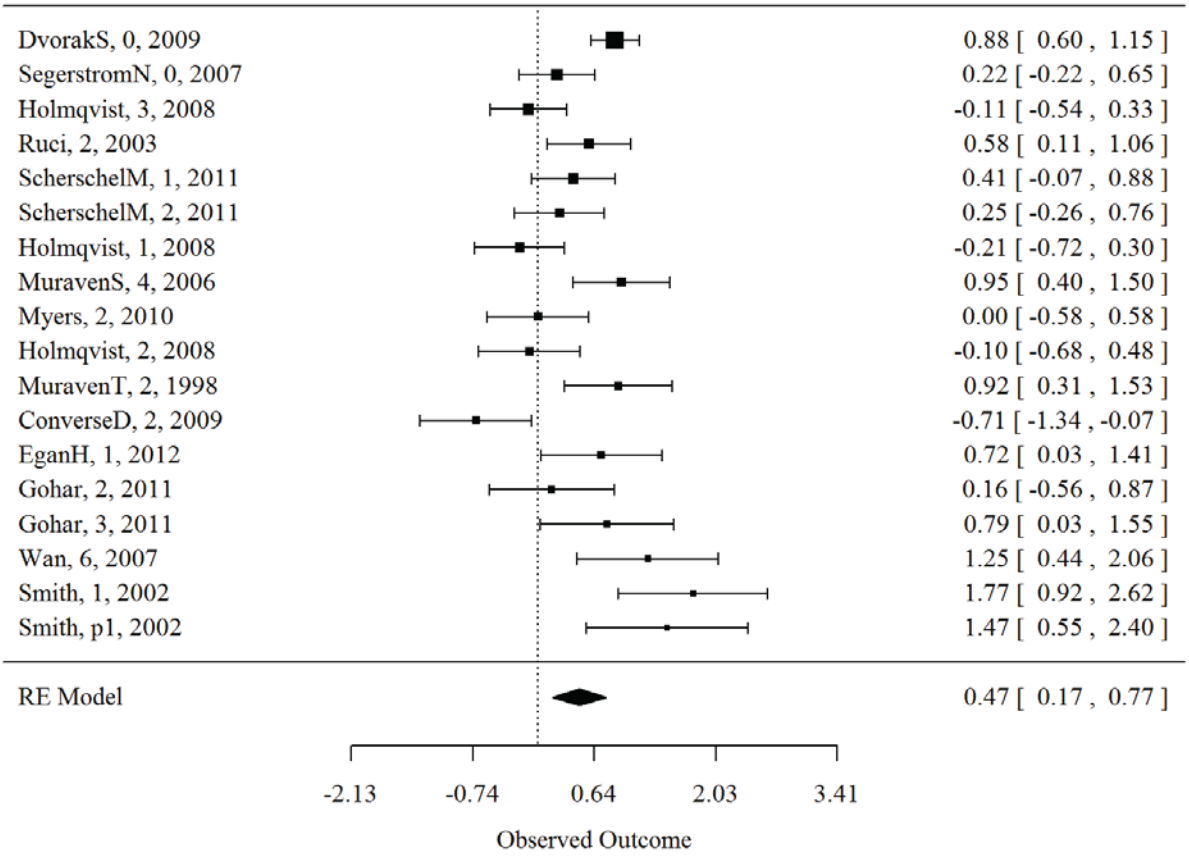
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 3. Forest plot of the experiments in the hand grip sample.



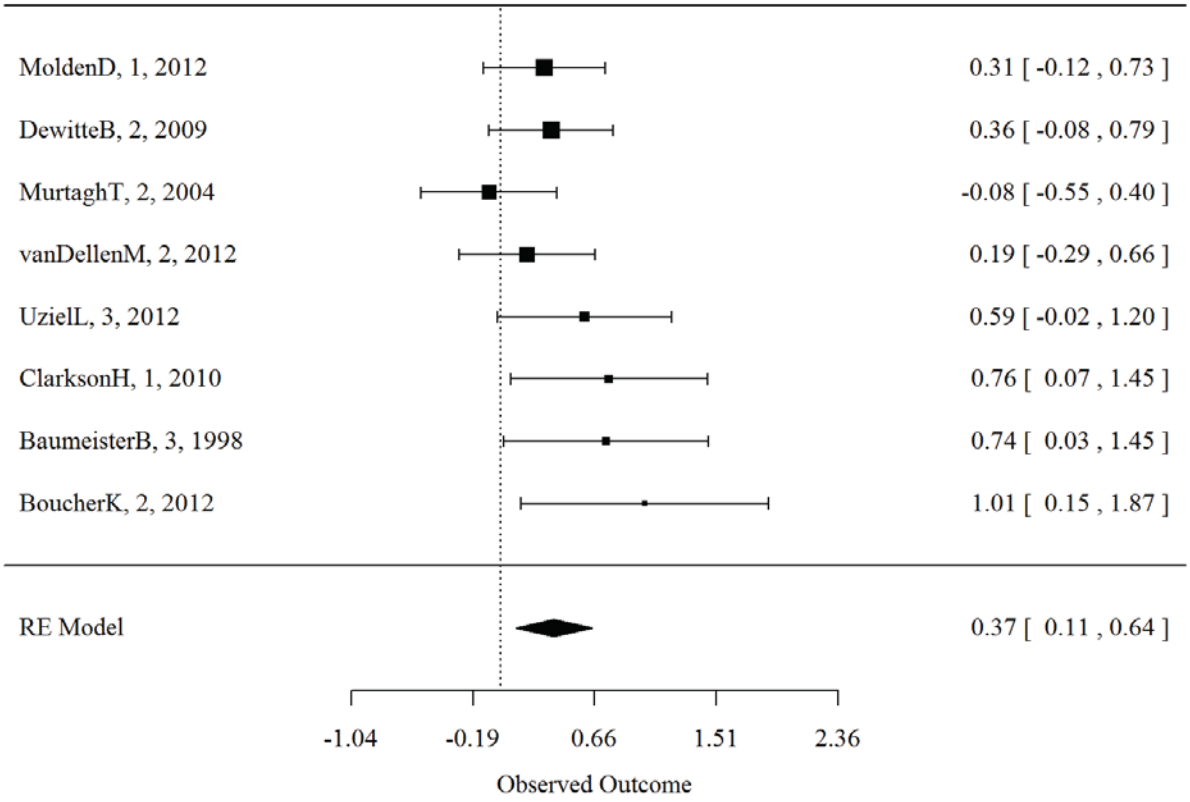
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 4. Forest plot of the experiments in the impossible anagrams sample.



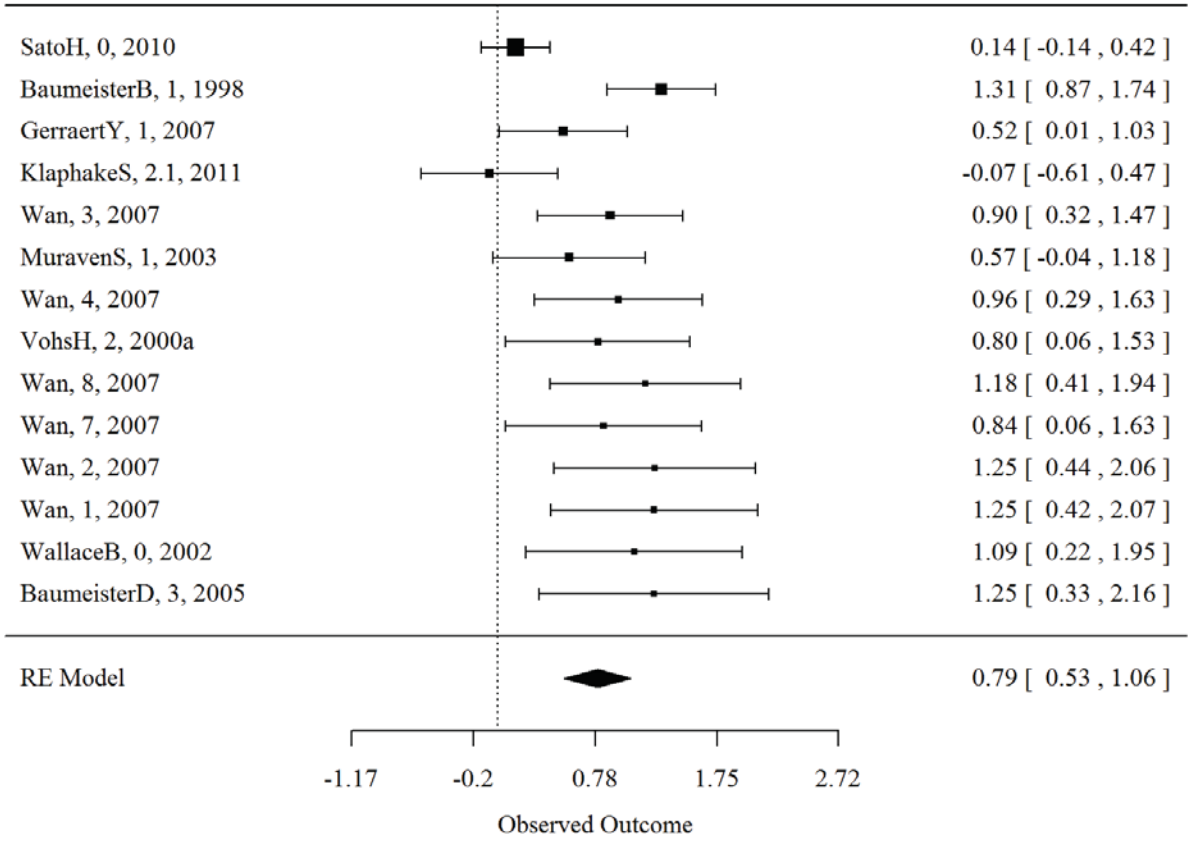
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 5. Forest plot of the experiments in the possible anagrams sample.



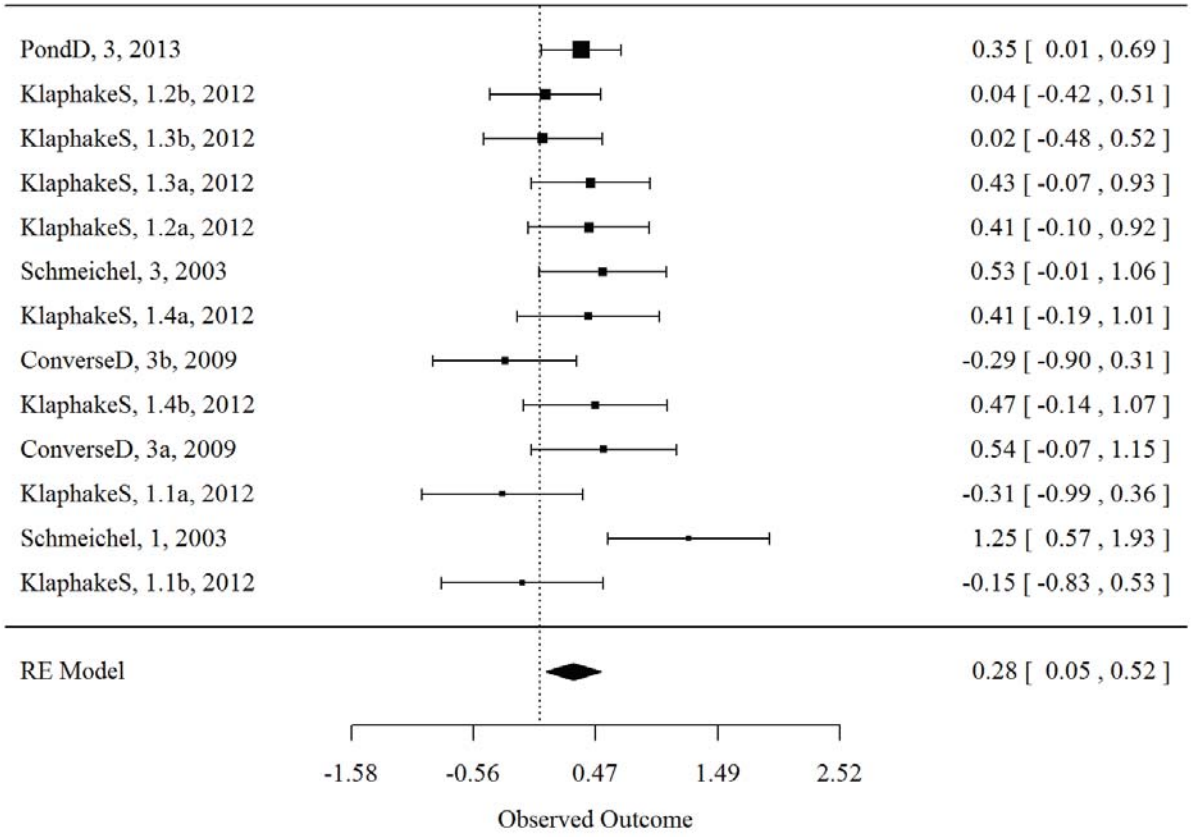
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 6. Forest plot of the experiments in the impossible puzzles sample.



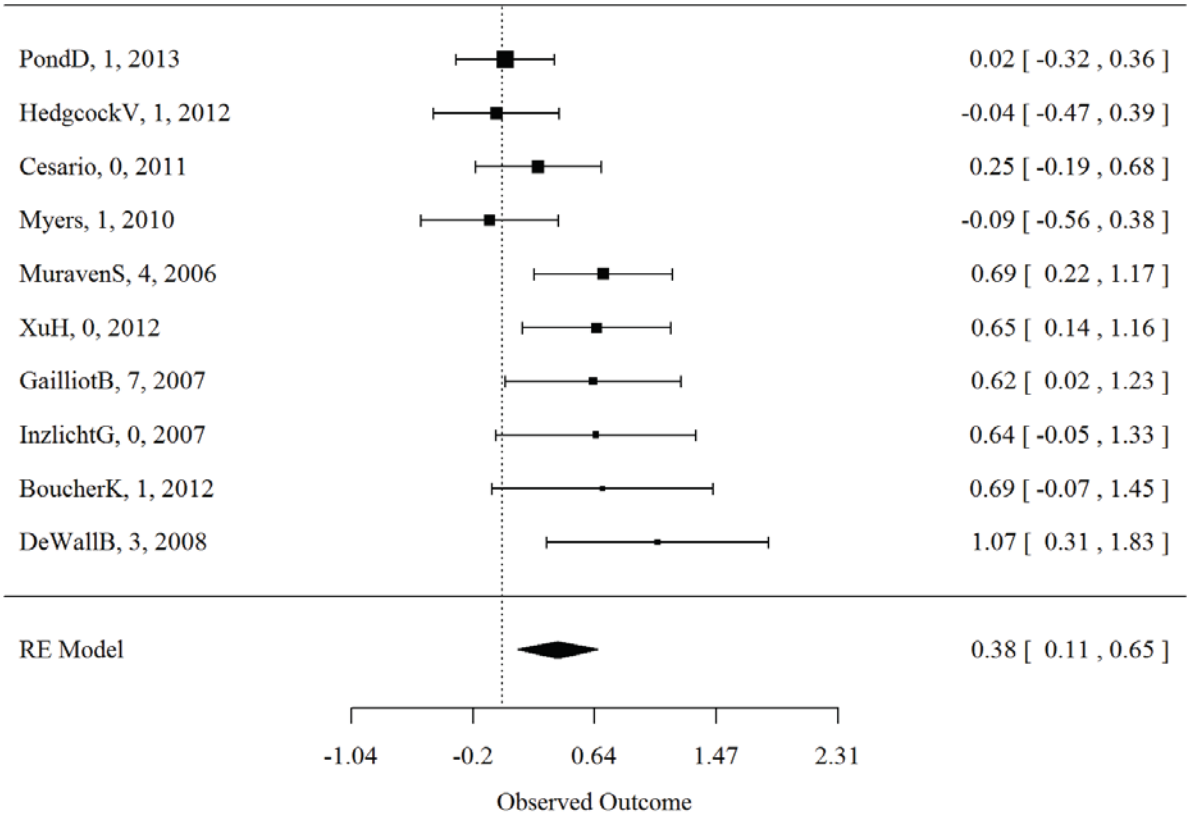
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 7. Forest plot of the experiments in the standardized tests sample.



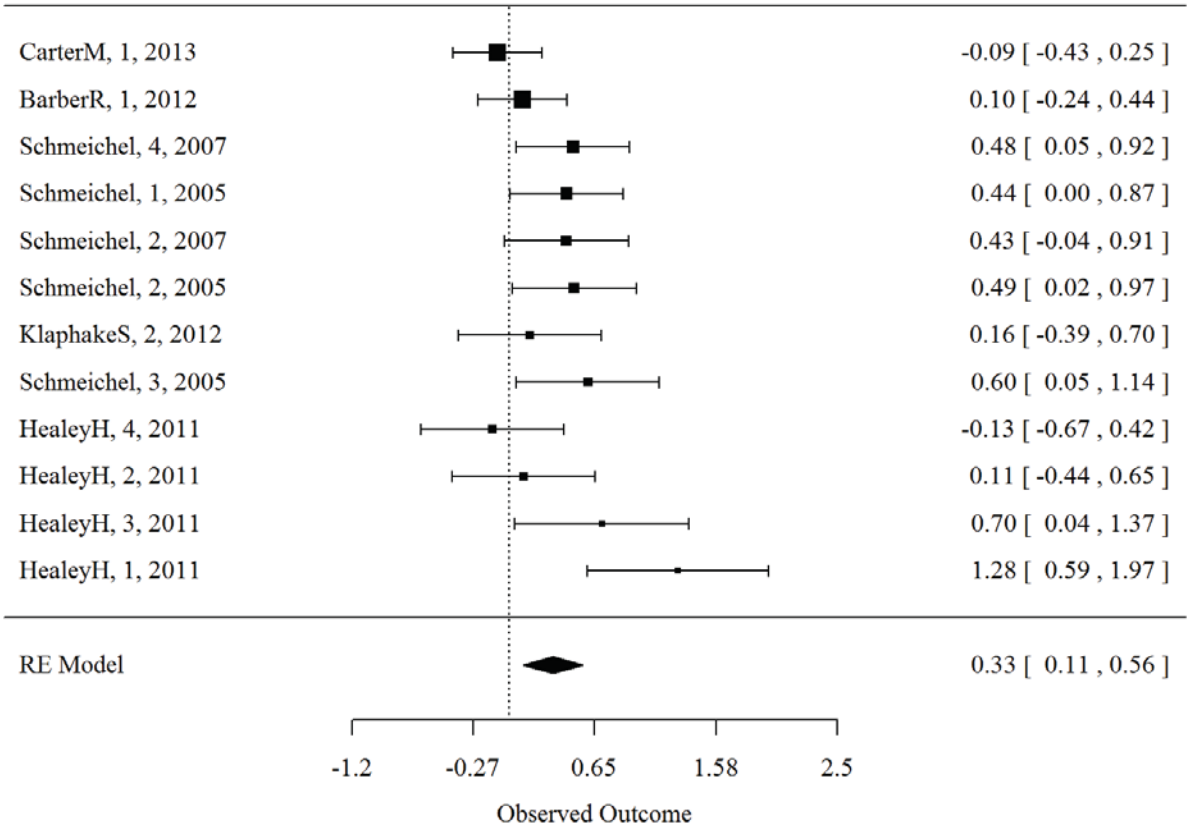
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 8. Forest plot of the experiments in the Stroop sample.



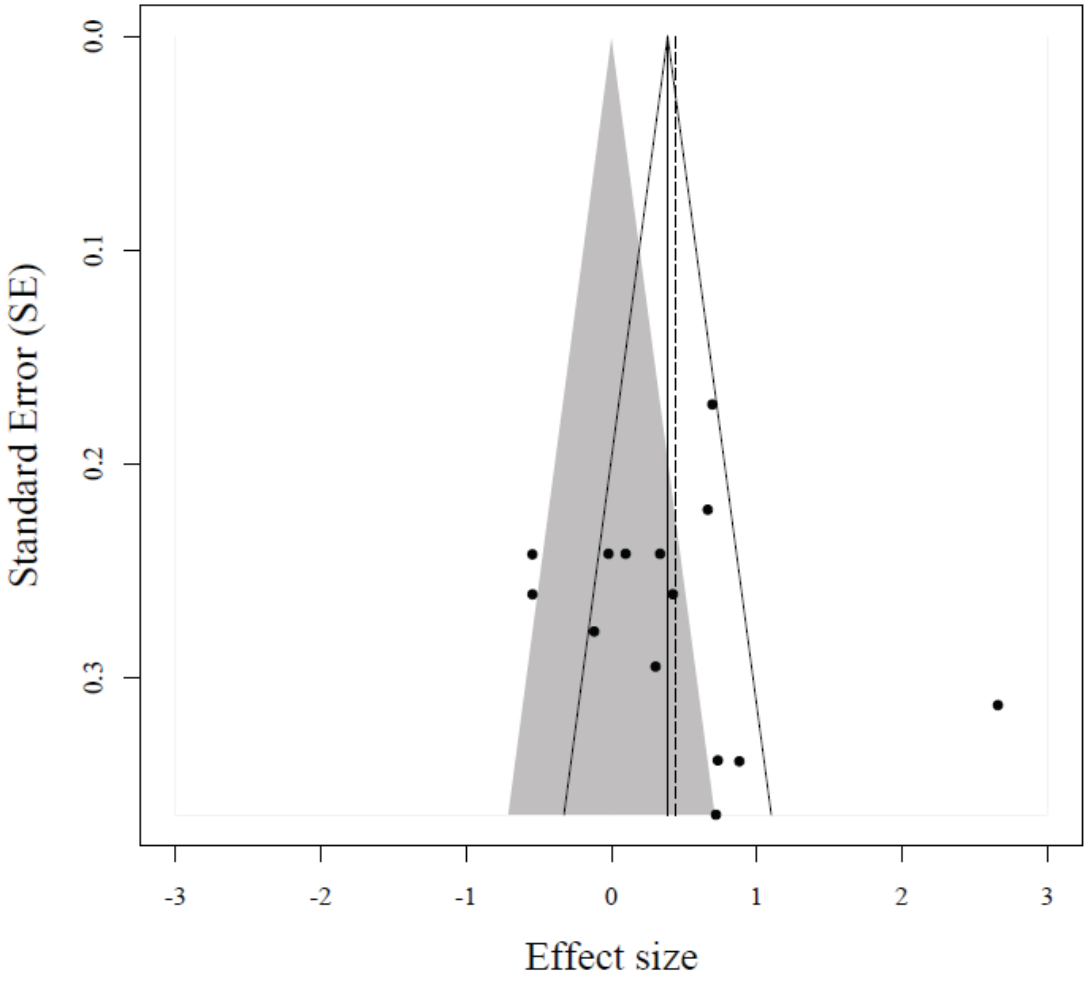
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 9. Forest plot of the experiments in the working memory sample.



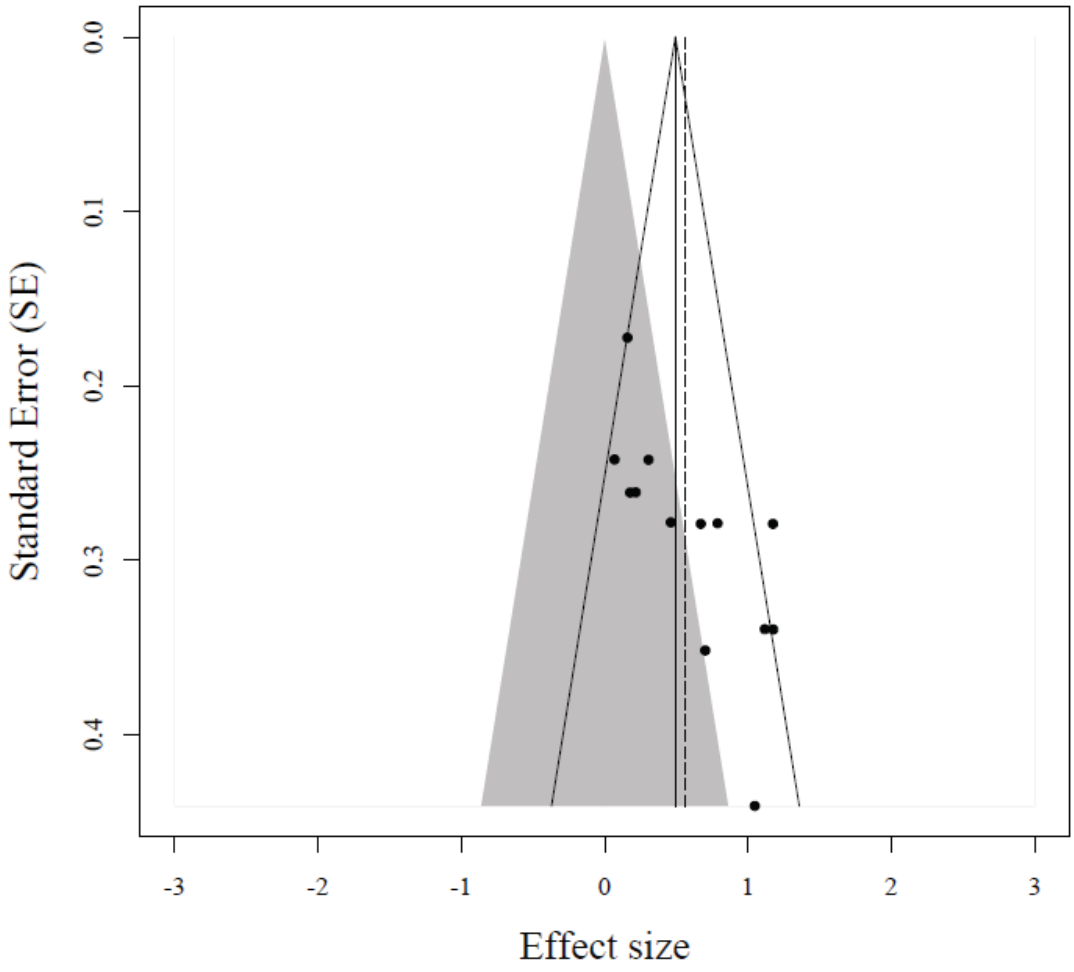
Note. The experiments are ordered by precision (i.e., the inverse of the standard error). The size of the black boxes for each experiment represents precision; the whiskers represent the 95% confidence intervals around the experiment-specific estimates. The dashed vertical line is placed at zero.

Figure 10. Contour-enhanced funnel plot of the experiments in the food consumption sample.



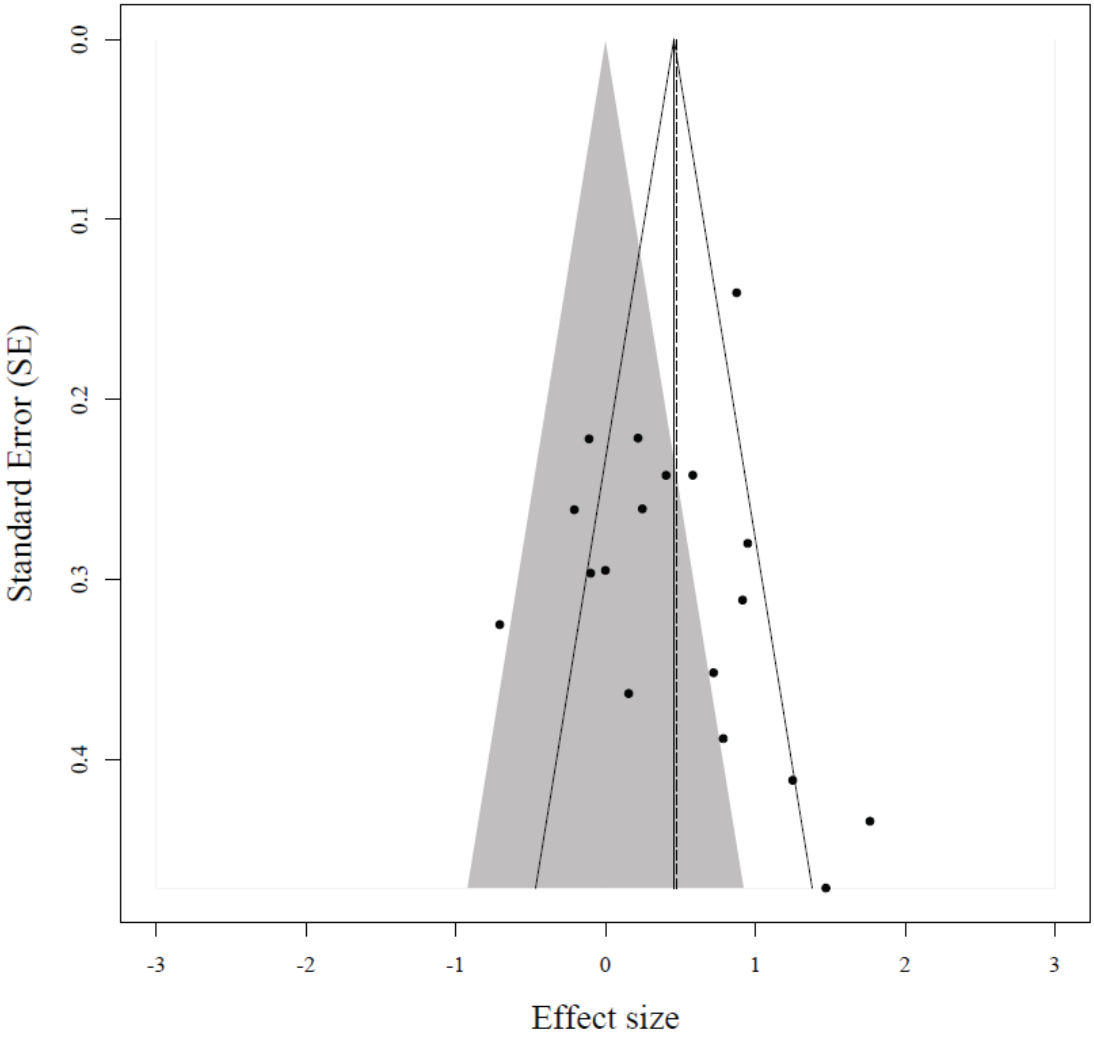
Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 11. Contour-enhanced funnel plot of the experiments in the hand grip sample.



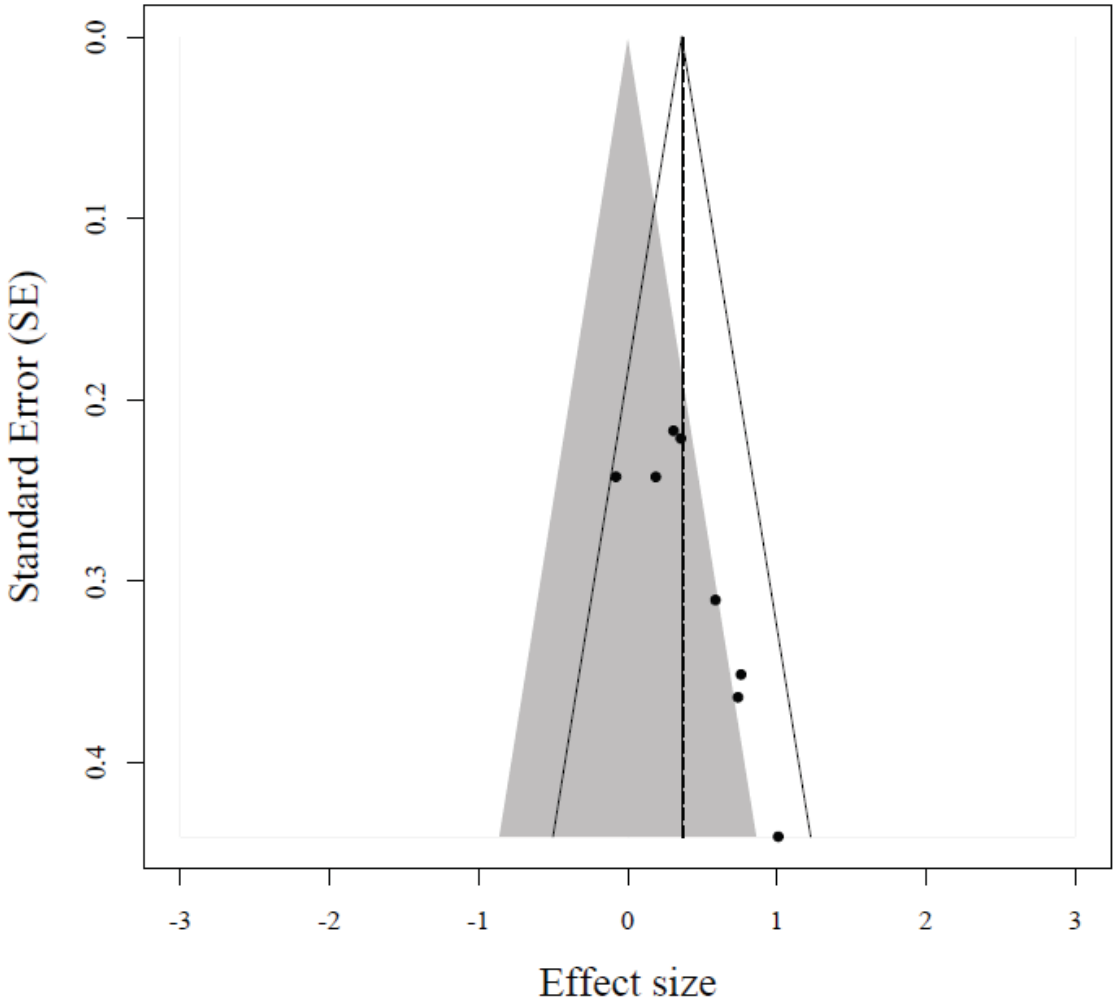
Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 12. Contour-enhanced funnel plot of the experiments in the impossible anagrams sample.



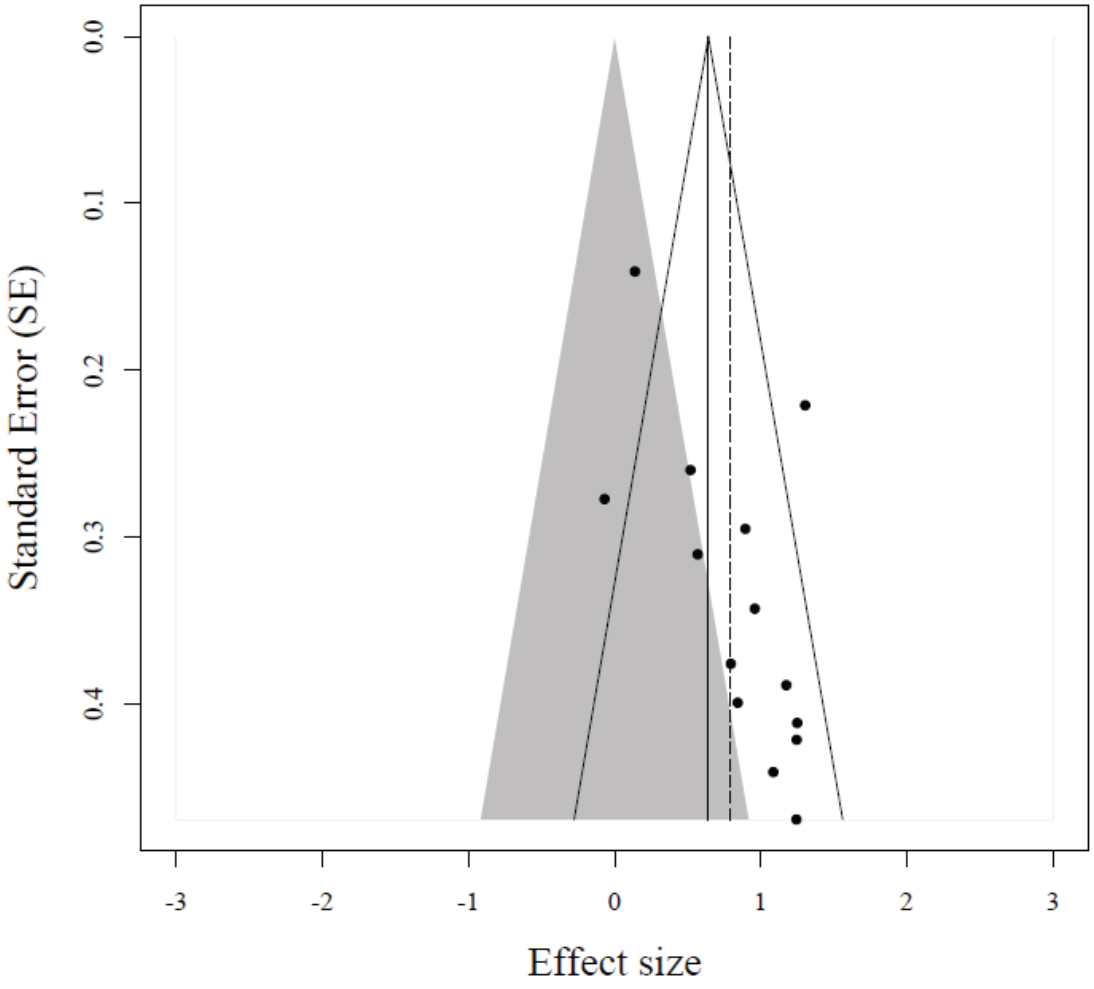
Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 13. Contour-enhanced funnel plot of the experiments in the possible anagrams sample.



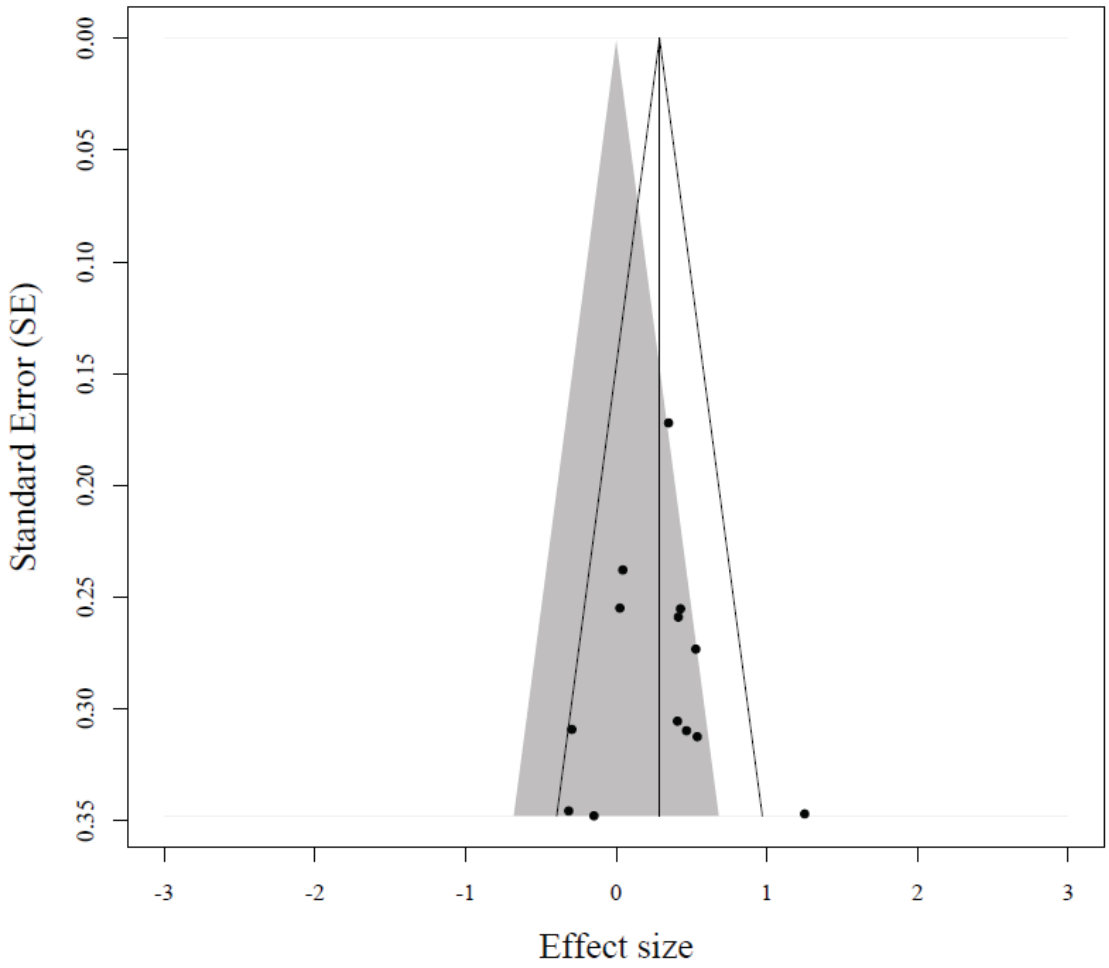
Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 14. Contour-enhanced funnel plot of the experiments in the impossible puzzles sample.



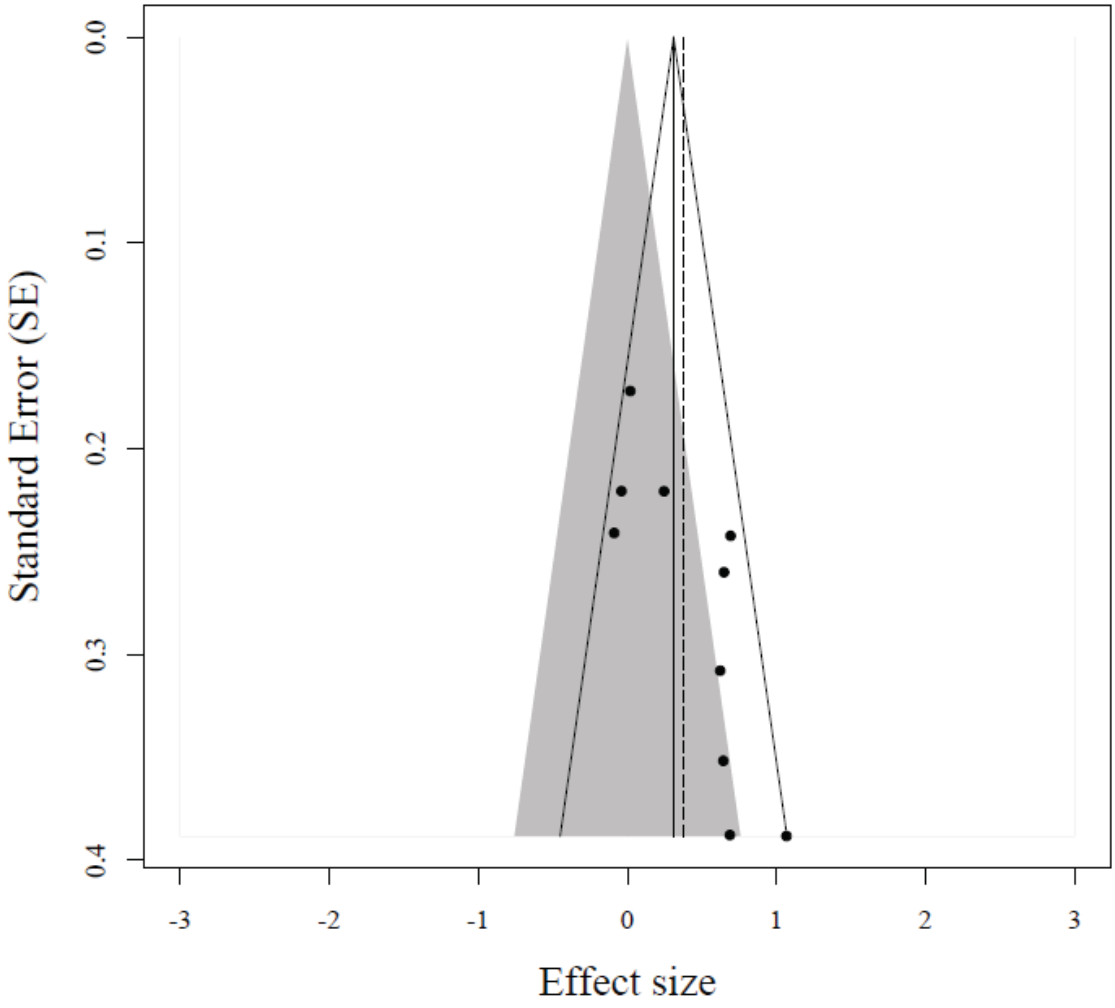
Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 15. Contour-enhanced funnel plot of the experiments in the standardized tests sample.



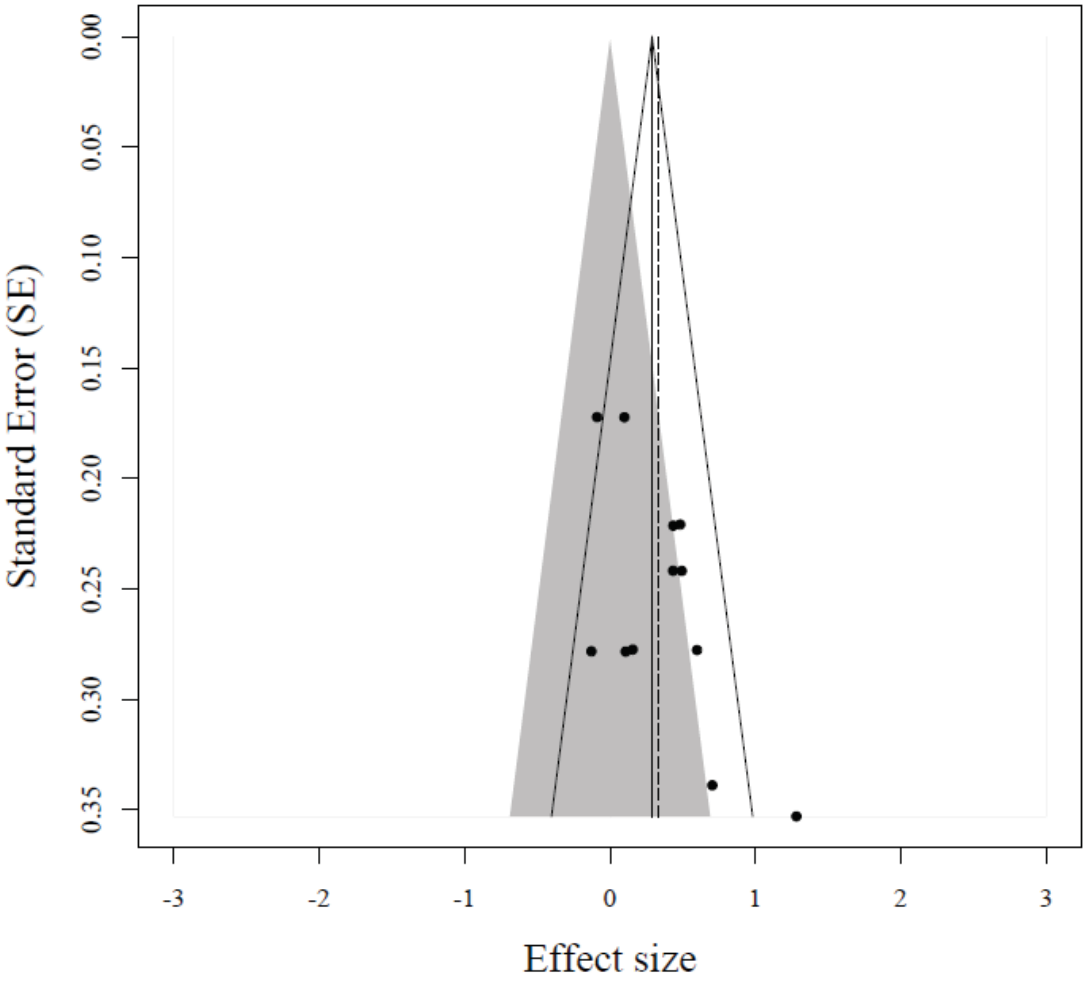
Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 16. Contour-enhanced funnel plot of the experiments in the Stroop sample.



Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Figure 17. Contour-enhanced funnel plot of the experiments in the working memory sample.



Note. Dots represent effect size estimates from individual experiments. Dots in the grey area represent nonsignificant results. The solid angled lines represent the confidence intervals for the fixed-effect model. The solid vertical line represents the estimate for the overall effect from the fixed-effect model. The dashed vertical line represents the estimate of the overall effect from the random-effects model.

Table 1. Coded characteristics of experiments using food consumption as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
BaumeisterD	2	2005	SE	0	0	1	1	0.88	0.12	19, 19
ChristiansenC	0	2012	EV	0	0	1	0	0.66	0.05	40, 40
DingemansM	0	2009	EV	0	0	1	0	-0.02	0.06	33, 33
FrieseH	2	2008	EV	0	0	1	0	0.34	0.06	33, 33
FrieseH	3	2008	EV	0	0	1	0	0.3	0.09	25, 21
HofmannR	0	2007	EV	0	0	1	0	-0.11	0.08	26, 24
ImhoffS	1	2013	S	0	0	1	0	0.69	0.03	69, 68
MuravenC	0	2002	TS	0	0	1	1	0.42	0.07	29, 29
OatenW	1	2008	SE	0	0	1	0	2.66	0.09	37, 36
StillmanT	3	2009	AV	0	0	1	1	0.09	0.06	33, 33
VohsH	3	2000	EV	0	0	1	1	0.73	0.11	18, 18
LattimoreM	0	2004	S	0	0	1	0	-0.54	0.07	29, 30
VohsB	5	2013	AE	0	1	1	1	0.72	0.13	15, 15
DewitteB	2	2009	FT	0	1	1	0	-0.54	0.06	35, 38
				AR = 0.67		*κ = 1*	*κ = .84*	*r = .99*	*r = .97*	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 2. Coded characteristics of experiments using the hand grip as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
BrayA	0	2008	S	0	0	1	0	0.46	0.08	26, 23
BrayA	0	2011	S	0	0	1	0	0.18	0.07	33, 28
EganH	2	2012	CL	0	0	1	0	1.18	0.12	21, 20
Litvin	0	2012	TS	0	0	1	0	0.16	0.03	54, 108
MartijnT	1	2002	EV	0	0	1	0	0.7	0.12	17, 16
MoldenD	2	2012	CL	0	0	1	0	1.05	0.19	11, 11
MuravenT	1	1998	EV	0	0	1	1	0.67	0.08	40, 20
MurtaghT	1	2004	S	0	0	1	0	0.07	0.06	42, 27
Neale-Lorello	1	2009	CL	0	0	0	0	0.22	0.07	30, 29
SeeleyG	1	2003	TS	0	0	1	0	0.31	0.06	37, 36
SeeleyG	2	2003	TS	0	0	1	0	0.79	0.08	28, 27
TylerB	2	2009	CL	0	0	1	0	1.17	0.08	30, 30
TylerB	3	2009	TS	0	0	1	0	1.12	0.12	20, 20
				AR = 1		$\kappa = 1$	$\kappa = 1$	$r = .99$	$r = .99$	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 3. Coded characteristics of experiments using impossible anagrams as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
DvorakS	0	2009	EV	0	0	1	0	0.87	0.02	90, 90
EganH	1	2012	TS	0	0	1	0	0.72	0.12	17, 16
Gohar	2	2011	T	0	0	0	0	0.16	0.13	14, 14
Gohar	3	2011	T	0	0	0	0	0.79	0.15	16, 12
MuravenT	2	1998	TS	0	0	1	1	0.92	0.09	17, 34
Myers	2	2010	AV	0	0	0	0	0	0.09	25, 21
ScherschelM	1	2011	AE	0	0	0	0	0.41	0.06	35, 33
ScherschelM	2	2011	T	0	0	0	0	0.25	0.07	24, 31
SegerstromN	0	2007	FT	0	0	1	0	0.22	0.05	41, 42
Smith	1	2002	TS	0	0	0	1	1.77	0.19	14, 14
Smith	p1	2002	TS	0	0	0	1	1.47	0.22	10, 12
Wan	6	2007	TS	0	0	0	0	1.25	0.17	14, 13
Holmqvist	1	2008	AV	0	0	0	0	-0.21	0.07	33, 29
Holmqvist	2	2008	WM + AV	1	0	0	0	-0.09	0.09	51, 15
Holmqvist	3	2008	WM + AV	1	0	0	0	-0.11	0.05	74, 27
ConverseD	2	2009	CL + S	1	0	1	0	-0.71	0.12	20, 20
MuravenS	4	2006	T	0	1	1	1	0.95	0.08	57, 19
Ruci	2	2003	S	0	1	0	0	0.58	0.06	30, 37
				AR = 1		$\kappa = 1$	$\kappa = 1$	$r = .99$	$r = .93$	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 4. Coded characteristics of experiments using possible anagrams as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
BaumeisterB	3	1998	EV	0	0	1	1	0.74	0.13	15, 15
BoucherK	2	2012	TS	0	0	1	0	1.01	0.19	11, 11
ClarksonH	1	2010	CL	0	0	1	0	0.76	0.12	16, 16
MoldenD	1	2012	CL	0	0	1	0	0.31	0.05	43, 42
MurtaghT	2	2004	TS	0	0	1	0	-0.08	0.06	26, 50
vanDellenM	2	2012	AV	0	0	1	0	0.19	0.06	56, 22
UzielL	3	2012	T	0	0	1	1	0.59	0.09	20, 23
DewitteB	2	2009	FT	0	1	1	0	0.36	0.05	38, 38
				AR = 1		$\kappa = 1$	$\kappa = .6$	$r = .99$	$r = .89$	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 5. Coded characteristics of experiments using impossible puzzles as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
BaumeisterB	1	1998	FT	0	0	1	1	1.31	0.05	25, 42
BaumeisterD	3	2005	SE	0	0	1	1	1.25	0.22	10, 10
KlaphakeS	2	2011	AE	0	0	0	1	-0.07	0.08	20, 20
MuravenS	1	2003	TS	0	0	1	1	0.57	0.09	22, 21
SatoH	0	2010	CL	0	0	1	0	0.14	0.02	86, 109
VohsH	2	2000	FT	0	0	1	1	0.79	0.14	14, 14
WallaceB	0	2002	S	0	0	1	1	1.09	0.19	11, 11
Wan	1	2007	CL	0	0	0	0	1.25	0.18	13, 12
Wan	2	2007	CL	0	0	0	0	1.25	0.17	14, 13
Wan	3	2007	CL	0	0	0	0	0.89	0.09	24, 24
Wan	4	2007	CL	0	0	0	0	0.96	0.12	39, 38
Wan	7	2007	CL	0	0	0	0	0.84	0.16	13, 13
Wan	8	2007	CL	0	0	0	0	1.18	0.15	15, 14
GerraertY	1	2007	FT	0	0	1	0	0.52	0.07	24, 20
				AR = 1		$\kappa = 1$	$\kappa = .88$	$r = .95$	$r = .97$	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 6. Coded characteristics of experiments using standardized tests as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
KlaphakeS	1.1a	2012	AE	0	0	0	1	-0.31	0.12	10, 10
KlaphakeS	1.1b	2012	WM	0	0	0	1	-0.15	0.12	11, 10
KlaphakeS	1.2a	2012	AE	0	0	0	1	0.41	0.07	19, 20
KlaphakeS	1.2b	2012	WM	0	0	0	1	0.04	0.06	20, 26
KlaphakeS	1.3a	2012	AE	0	0	0	1	0.43	0.07	20, 20
KlaphakeS	1.3b	2012	WM	0	0	0	1	0.02	0.07	20, 20
KlaphakeS	1.4a	2012	AE	0	0	0	1	0.41	0.09	13, 14
KlaphakeS	1.4b	2012	WM	0	0	0	1	0.47	0.12	14, 13
Schmeichel	1	2003	AV	0	0	1	1	1.25	0.07	12, 12
Schmeichel	3	2003	AV	0	1	1	1	0.53	0.09	18, 18
ConverseD	3a	2009	CL	0	0	1	1	0.54	0.09	15, 15
ConverseD	3b	2009	CL + S	1	0	1	1	-0.29	0.09	15, 15
PondD	3	2013	AV	0	0	0	1	0.35	0.03	65, 63
				AR = 0.50		$\kappa = 1$	$\kappa = 1$	$r = .99$	$r = .98$	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 7. Coded characteristics of experiments using Stroop as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
BoucherK	1	2012	CL	0	0	1	0	0.69	0.15	14, 13
Cesario	0	2013	AV	0	0	0	0	0.25	0.05	31, 30
DeWallB	3	2008	SE	0	0	1	1	1.07	0.15	14, 14
GailliotB	7	2007	AV	0	0	1	1	0.62	0.09	16, 15
InzlichtG	0	2007	EV	0	0	1	1	0.64	0.12	15, 18
Myers	1	2010	T	0	0	0	0	-0.09	0.06	24, 26
PondD	1	2013	AV	0	0	0	1	0.02	0.03	56, 60
XuH	0	2012	EV	0	0	1	0	0.65	0.07	24, 23
HedgcockV	1	2012	AV	0	1	1	1	-0.04	0.05	30, 30
MuravenS	4	2006	T	0	1	1	1	0.69	0.06	38, 38
				AR = 1		$\kappa = 1$	$\kappa = .87$	$r = .90$	$r = .85$	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 8. Coded characteristics of experiments using working memory as an outcome task.

Author(s)	Exp	Year	IV	Mult. IV	Mult. DV	Pub	Lab	*g*	*v*	*n1, n2*
HealeyH	1	2011	AV	0	0	1	0	1.28	0.12	19, 19
HealeyH	2	2011	AV	0	0	1	0	0.12	0.08	25, 25
HealeyH	3	2011	AV	0	0	1	0	0.7	0.11	19, 18
HealeyH	4	2011	AV	0	0	1	0	-0.13	0.08	27, 22
Schmeichel	1	2005	AV	0	0	1	1	0.44	0.05	41, 38
Schmeichel	2	2005	AV	0	0	1	1	0.49	0.06	31, 31
Schmeichel	3	2005	EV	0	0	1	1	0.59	0.08	22, 22
Schmeichel	2	2007	AE	0	0	1	1	0.43	0.06	32, 29
Schmeichel	4	2007	ES	0	0	1	1	0.48	0.05	32, 33
BarberR	1	2012	AE	0	1	1	0	0.09	0.03	76, 76
KlaphakeS	2	2012	AE	0	1	0	1	0.16	0.08	21, 21
CarterM	1	2013	AE + AV	1	0	1	0	-0.09	0.03	71, 71
				AR = 1		κ = 1.00	κ = 1.00	*r* = .99	*r* = .99	

Note. Author names are given as the last name of the first author and the first letter of the last name of the second author. Exp is the number given to the experiment in the original paper (0 = only one experiment was conducted in the original paper). IV = the outcome task: AV = attention video; CL = crossing letters; TS = thought suppression; EV = emotion video; S = Stroop; AE = attention essay; FT = food temptation; T = transcription; SE = social exclusion; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister/Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* is the adjusted standardized mean difference and *v* is its associated variance. Inter-rater reliability is given as agreement rate (AR) for the three codes associated with the manipulation and outcome tasks, Cohen's κ for categorical variables, and Pearson's *r* for continuous variables.

Table 9. Random-effects models.

	Estimate of average overall effect			Heterogeneity			
Sample	*g**	*k*	*p*	*Q*	*p*	τ^2, τ	I^2 (95%CI)
Food Consumption	0.44 (-0.01, 0.89)	14	0.06	96.75	<0.001	0.52, 0.72	88.54% (79.1%-91.0%)
Hand Grip	0.56 (0.31, 0.81)	13	<0.001	26.85	0.008	0.09, 0.31	55.73% (16.7%, 76.0%)
Impossible Anagrams	0.47 (0.17, 0.77)	18	0.005	67.4	<0.001	0.25, 0.50	77.19% (60%, 84.1%
Possible Anagrams	0.37 (0.11, 0.64)	8	0.01	8.95	0.26	0.01, 0.11	13.8% (0%, 63.8%)
Impossible Puzzles	0.79 (0.53, 1.06)	14	<0.001	39.40	<0.001	0.14, 0.37	61.17% (42.2%, 81.2%)
Standardized Tests	0.28 (0.05, 0.52)	13	0.02	20.51	0.06	0.04, 0.21	36.57% (0%, 69.5%)
Stroop	0.38 (0.12, 0.65)	10	0.01	19.05	0.02	0.08, 0.28	53.54% (3.2%, 76.9%)
Working Memory	0.33 (0.11, 0.56)	12	0.008	21.9	0.03	0.05, 0.23	47.83% (2.5%, 74.18%)

Note. Numbers given in parentheses are the lower and upper limits of the 95% confidence intervals.

Table 10. Mixed-effects models.

	Model results			Test of moderators		Residual heterogeneity	
Sample	Moderator	*β*	*p*	*F*	*p*	*Qe*	*p*
Food consumption	Intercept	0.43	0.16	0.37	0.7	87.68	<.001
	Source Lab	0.23	0.64				
	Multiple DV	-0.48	0.47				
Impossible anagrams	Intercept	0.47	0.007	4.82	0.01	30.8	0.004
	Publication	-0.07	0.75				
	Source Lab	0.79	0.02				
	Multiple IV	-0.72	0.02				
	Multiple DV	-0.05	0.88				
Impossible puzzles	Intercept	0.85	0.001	0.09	0.91	34.25	<0.001
	Publication	-0.13	0.68				
	Source Lab	0.05	0.88				
Stroop	Intercept	0.06	0.73	3.43	0.09	7.16	0.31
	Publication	0.65	0.02				
	Source Lab	<0.01	0.98				
	Multiple DV	-0.41	0.15				
Working memory	Intercept	0.27	0.13	0.95	0.42	16.63	0.05
	Source Lab	0.20	0.36				
	Multiple DV	-0.22	0.42				

Note. Intercept corresponds to β_0 as discussed in the text. *Qe* = the *Q* statistic used to test for residual heterogeneity.

Table 11. IC-index and average power.

Sample	Measure	Ind. Exp.	RE-LL	RE	RE-UL
Food consumption	IC-index	0.65	0.99	0.86	<0.01
	Avg. Power	0.48	0.05	0.39	0.89
Hand Grip	IC-index	0.48	0.99	0.43	0.01
	Avg. Power	0.51	0.21	0.53	0.8
Impossible Anagrams	IC-index	0.61	0.99	0.9	0.05
	Avg. Power	0.49	0.1	0.38	0.81
Possible Anagrams	IC-index	0.39	0.99	0.66	0.06
	Avg. Power	0.37	0.07	0.26	0.59
Impossible Puzzles	IC-index	0.80	0.99	0.83	0.25
	Avg. Power	0.64	0.39	0.63	0.81
Standardized Tests	IC-index	0.13	0.85	0.45	0.04
	Avg. Power	0.87	0.15	0.55	0.96
Stroop	IC-index	0.45	0.99	0.73	0.06
	Avg. Power	0.38	0.07	0.27	0.59
Working Memory	IC-index	0.78	0.99	0.93	0.21
	Avg. Power	0.35	0.07	0.27	0.58

Note. Values given for Ind. Exp. Are the IC-index and average power calculated based on the effect estimated provided by the individual experiments. Values given for RE-LL, RE, and RE-UL are the IC-index and average power calculated based on the lower-limit of the confidence interval surrounding the estimate from the RE model, the estimate from the RE model, and the upper limit of the confidence interval, respectively.

Table 12. Regression-based methods.

	PET		PEESE	
Sample	b_0	b_1	b_0	b_1
Food Consumption	-0.21 (-2.35, 1.93)	2.38 (-5.93, 10.69)	-0.01 (-1.13, 1.11)	6.0 (-9.73, 21.73)
Hand Grip	-0.76 (-1.55, 0.04)	4.76 (1.81, 7.71)	-0.11 (-0.05, 0.32)	8.32 (2.96, 13.69)
Impossible Anagrams	0.39 (-0.53, 1.30)	0.26 (-3.13, 3.66)	0.31 (-0.20, 0.83)	1.99 (-4.03, 8.0)
Possible Anagrams	-0.62 (-1.38, 0.13)	3.70 (0.93, 6.46)	-0.09 (-0.46, 0.28)	6.04 (1.68, 10.39)
Impossible Puzzles	-0.16 (-0.80, 0.48)	3.01 (0.77, 5.24)	0.22 (-0.19, 0.63)	5.11 (1.12, 9.23)
Standardized Tests	0.33 (-0.76, 1.42)	-0.17 (-4.19, 3.85)	0.31 (-0.31, 0.92)	-0.27 (-8.08, 7.53)
Stroop	-0.72 (-1.48, 0.03)	4.21 (1.23, 7.19)	-0.17 (-0.57, 0.23)	7.44 (1.95, 12.93)
Working Memory	-0.61 (-1.42, 0.20)	3.87 (0.49, 7.26)	-0.16 (-0.59, 0.26)	7.94 (1.13, 14.74)

Note. b_0 represents the intercept, or the estimate of the average overall effect adjusted for small-study effects, and b_1 represents the slope, or the test for funnel plot asymmetry. Numbers given in parentheses are the lower and upper limits of the 95% confidence intervals.

Table 13. Trim and fill.

Sample	*g**	*+k*	*p*
Food consumption	0.44 (-0.01, 0.89)	0	0.06
Hand Grip	0.36 (0.09, 0.63)	4	0.008
Impossible Anagrams	0.41 (0.13, 0.70)	1	0.005
Possible Anagrams	0.26 (0.04, 0.48)	3	0.022
Impossible Puzzles	0.63 (0.38, 0.88)	4	<0.001
Standardized Tests	0.28 (0.05, 0.52)	1	0.02
Stroop	0.17 (-0.09, 0.43)	4	0.21
Working Memory	0.33 (0.11, 0.56)	0	0.008

Note. +k is the number of experiments imputed by the trim and fill. Numbers given in parentheses are the lower and upper limits of the 95% confidence intervals.

Table 14. Interpreting the results in terms of evidence for the depletion effect as laid out in the limited strength model of self-control.

Key Questions	Sample								Basis for the conclusion	
	FC	HG	IA	PA	IP	ST	S	WM	Test	Table
Q1: Is the average overall depletion effect statistically significant?	N	Y	Y	Y	Y	Y	Y	Y	RE model	9
Q2: After imputing experiments that are potentially missing due to publication bias, is the overall average depletion effect still significant?	.	Y	Y	Y	Y	Y	N	Y	Trim & fill	13
Q3: Is the overall average depletion effect still significant after correcting for small-study effects?	.	N	.	N	N	.	N	N	PET & PEESE	12
Q4: When the overall average depletion effect is moderated by an observed experiment-level characteristics, is the effect significant at all levels of the moderator(s)?	.	.	N	.	.	.	N	.	ME model	10
Q5: Does the evidence support the existence of the depletion effect as proposed in the limited strength model (i.e., the answers to Q1 through Q4 are all "Y")?	N	N	N	N	N	Y	N	N		

Note. FC = food consumption sample; HG = handgrip sample; IA = impossible anagram sample; PA = possible anagram sample; IP = impossible puzzles sample; ST = standardized tests sample; S = Stroop sample; WM = working memory sample. "Y" = yes; "N" = no; "." = not applicable. Random-effects model = the RE model; Mixed-effects mode = the ME model.

Endnotes

[1] Hagger et al. included only the effect size estimate from the "control" levels of experimentally manipulated moderators in an attempt to derive "a simple, unattenuated test of the ego-depletion effect" (p. 504). For example, in Gailliot et al. (2007), participants were either given glucose or an artificial sweetener control in addition to the standard sequential task paradigm. Hagger et al. included only the effect of the depletion manipulation for those participants who received the artificial sweetener control. It is possible that this was the correct approach; however, the appropriateness of this decision depends entirely on whether the moderator truly affects the depletion effect in the way that it is hypothesized. In other words, if glucose administration has no effect on depletion, then not including the effect size derived from the performance of participants who received glucose biases the overall estimate of the effect in favor of confirming the existence of the depletion effect. To correctly account for this, one would have to make the decision on a case-by-case basis, and such decisions would need to be based on (sometimes very limited) previous research. Hagger et al.'s method is attractive because it lessens the number of relatively subjective judgment calls required by the meta-analyst, though any estimate of the true effect size resulting from such a synthesis will almost certainly tend to favor the depletion hypothesis.

[2]When sample size is small, *d* has a slight tendency to overestimate δ. According to Hedges (1981), this bias can be corrected through the use of a correction factor, *J*, resulting in a bias-correct estimate known as Hedges' g. The correction factor is calculated as

$$J(df) = 1 - \frac{3}{4df-1},$$

where *df* is denominator for S_{within}, $n_1 + n_2 - 2$, which is the degrees of freedom when estimating the pooled standard deviation. Hedges' g is

$$g = J(df)d,$$

and the variance of g is

$$v_g = [J(df)]^2 v_d.$$

For both *d* and *g*, the standard error is the square root of the variance.

[3] For this dissertation, the reference distribution used in significance testing for point estimates is a *t*-distribution because the Knapp and Hartung adjustment (Knapp & Hartung, 2003) was applied.

Appendix A

The following online databases were searched: ISI Web of Science, EBSCO (including MEDLINE, PsychINFO, PsychARTICLES, PsychEXTRAS, ERIC), and ProQuest (including American Periodicals, Ethnic NewsWatch, FRANCIS, GenderWatch, PAIS, PILOTS, ProQuest Dissertations & Theses: History, ProQuest Dissertations & Theses: Social Sciences, ProQuest Research Library: Social Sciences, ProQuest Social Science Journals, ProQuest Sociology, Social Services Abstracts, Sociological Abstracts). Publication type was set to articles, proceedings papers, reviews, and meeting abstracts for ISI Web of Science; periodicals, reviews, reports, and dissertations for EBSCO; and conference papers and proceedings, dissertations and theses, reports, and scholarly journals for ProQuest. Each search was limited to results in English that used human subjects and that were dated from 1998 to 2012.

Exact search terms were as follows (an asterisk indicates a truncated search word, which includes all versions of the word in the search; for example, “deplet*” includes the words deplete, depletion, depleted, and depletes in the search): For ISI Web of Science, the full search term was (“Self Regulat*” or “Self Control” or “Impulse” or “Ego”) AND (“Resource” or “Deplet*” or Perform*”). For EBSCO, the full search term was (“Self Regulat*” or “Self Control” or “Impulse” or “Ego”) AND (“Resource” or “Deplet*” or Perform*”). And for ProQuest, the full search term was ((EXACT ("Self Control" or "self regulat*")) OR ("implus*" or "ego")) AND ((deplete* or resource* or perform*)) AND CAU(Baumeister R). In the search term for ProQuest, the code CAU(Baumeister R) specifies that the search only return hits that cite an author with the last name Baumeister

and first initial R. This option was only available for ProQuest, but reduced the total returned hits by several thousand.

Each search returned the following number of hits: 3,851 for ISI Web of Science, 7,889 for EBSCO, and 853 for ProQuest. These abstracts were then examined for general relevance. This resulted in 177 abstracts for ISI Web of Science, 132 abstracts for EBSCO, and 54 for ProQuest. With duplicates removed, this resulted in a combined total of 287 abstracts. From this list of abstracts, I obtained 269 full-text articles. From these articles, there were 328 independent experiments within 141 articles that made use of the sequential task paradigm.

Conference programs for the annual meetings of Society for Personality and Social Psychology (SPSP) and Association for Psychological Science (APS) were obtained for each year between 2003 (the earliest available year) and 2011. Using the find function, the search term "deplet" returned 31 poster and symposium presentation abstracts from the APS Convention Programs and 149 from the SPSP Meeting Programs. The authors for each of these posters or presentations were sent an email request for information about methods, statistics, and any other unpublished data.

In December 2012, a second wave of data collection was conducted to keep the dataset updated. This second wave was conducted in exactly the same way as the first, except that databases were searched from 2011 onward. From the online databases, each search returned the following number of hits: 1,209 for ISI Web of Science, 694 for EBSCO, and 72 for ProQuest. These abstracts were then examined for general relevance. This resulted in 90 abstracts for ISI Web of Science, 87 abstracts for EBSCO, and 14 for ProQuest. Removing duplicates yielded a combined total of 138 abstracts. From this list

of abstracts, I obtained 133 full-text articles. From these articles, there were 83 independent experiments within 47 articles that made use of the sequential task paradigm.

Conference programs for the annual meetings of Society for Personality and Social Psychology (SPSP) and Association for Psychological Science (APS) were searched for the years 2012 - 2013. Using the find function, the search term "deplet" returned 16 poster and symposium presentation abstracts from the APS Convention Programs and 54 from the SPSP Meeting Programs. The authors for each of these posters or presentations were sent the same email request for information.

In total, after adding experiments that were emailed to me to the set of experiments located via searching online databases, this search resulted in 498 individual instances of the sequential task paradigm. Each of these was then grouped by the type of manipulation task and the type of outcome task used. Following this, the dataset was organized by manipulation task in ascending order of the number of times each task was used. Ten categories of manipulation task emerged as frequently occurring (i.e., appearing ten or more times in the dataset). These ten categories were made up of a total of 322 experiments. The classes of manipulation task are listed in Table A1, along with the number of experiments in each. The 322 experiments that made use of a frequently used manipulation task were then organized by the type of outcome task used. The result was eight classes of outcome tasks that contained ten or more experiments (see Table A1). In total, my literature search produced 130 experiments that contained both frequently used manipulation tasks and frequently used outcome tasks. The categories of tasks are described in detail in the main text.

Table A1. Categories of frequently occurring manipulation and outcome tasks and the number of experiments that fall into them.

Manipulation task	*k*	Outcome task	*k*
Attention video	54	Impossible anagrams	24
Crossing letters	54	Hand grip	17
Thought suppression	42	Impossible puzzles	17
Emotion video	35	Food consumption	17
Stroop	40	Stroop	17
Attention essay	34	Possible anagrams	12
Food temptation	21	Standardized tests	14
Transcription	17	Working memory	14
Social exclusion	14		
Working memory	11		
Total	322	Total	130

Note. Two experiments are counted twice because they both occur in multiple samples (see main text).