Is Ego Depletion Real? An Analysis of Arguments

Malte Friese
Saarland University

David D. Loschelder
Leuphana University of Luneburg

Karolin Gieseler
Saarland University

Julius Frankenbach
Saarland University

Michael Inzlicht
University of Toronto

in press, *Personality and Social Psychology Review*
 The DOI will be provided as soon as it becomes available.

Abstract

An influential line of research suggests that initial bouts of self-control increase the susceptibility to self-control failure (ego depletion effect). Despite seemingly abundant evidence, some researchers have suggested that evidence for ego depletion was the sole result of publication bias and *p*-hacking, with the true effect being indistinguishable from zero. Here, we examine (a) whether the evidence brought forward against ego depletion will convince a proponent that ego depletion *does not* exist, and (b) whether arguments that could be brought forward in defense of ego depletion will convince a skeptic that ego depletion *does* exist. We conclude that despite several hundred published studies, the available evidence is inconclusive. Both, additional empirical and theoretical work is needed to make a compelling case for either side of the debate. We discuss necessary steps for future work toward this aim.

Keywords: ego depletion, self-control, self-regulation, replicability crisis, *p*-hacking

Author Note

Malte Friese, Department of Psychology, Saarland University; David D. Loschelder, Faculty of Business and Economics, Leuphana University of Lueneburg; Karolin Gieseler, Department of Psychology, Saarland University; Julius Frankenbach, Department of Psychology, Saarland University; Michael Inzlicht, Department of Psychology, Rotman School of Management, University of Toronto.

We thank Helena Kilger for her help in estimating the number of published depletion studies. We also thank Klaus Fiedler, two anonymous reviewers, and the handling editor Bertram Gawronski for their stimulating comments and excellent advice on how to improve this manuscript.

This work was supported by grants from the German Research Foundation to Malte Friese, Veronika Job, and David D. Loschelder (FR 3605/2-1) and to Malte Friese (FR 3605/3-1) as well as a grant from the Social Sciences and Humanities Research Council of Canada to Michael Inzlicht (435-2014-0556).

Correspondence concerning this article should be addressed to Malte Friese, Department of Psychology, Saarland University, Campus A2 4, 66123 Saarbrucken, Germany. Email: malte.friese@uni-saarland.de

Is Ego Depletion Real? An Analysis of Arguments

Imagine you are living in a capacious house with many other people. You feel comfortable and the house appears to be in good shape. One day, some of your housemates discover cracks in the floors of some of the rooms. "That's odd, this is a fairly new house", you think, and plenty of your friends have attested that the house is in good shape. Soon after, people start to systematically look for shortcomings, and discover major problems with the state of the house. Some people want to move out of the house right away; others are still torn and demand more information before making a decision.

The field of ego depletion research—the house in our metaphor—was thought to be in good shape not long ago. It was in such good shape, in fact, that it inspired research across the psychological disciplines and became one of the most famously discussed recent psychological discoveries in the public media, with even former President Obama claiming to have changed his behavior to not fall prey to it (Lewis, 2012). Ego depletion research faces serious threats today, however. Many scholars vehemently argue that the house should be abandoned—the sooner the better. Others are more reluctant, but certainly many are in doubt about the robustness of the house. How reliable and reproducible is the ego depletion effect? Has the effect been investigated in a less than optimal way all these years; or does it not even exist, after all?

In the present article, we consider what can be said for and what against the existence of ego depletion. Researchers on either side of the debate have expressed with great certainty that ego depletion does (not) exist. Is there conclusive evidence to warrant either claim? We strive to do justice to diverse viewpoints by reviewing and developing arguments both for and against the existence of ego depletion. On one hand, doubts about ego depletion have been so plentiful that a careful analysis of the evidence seems necessary before investing further financial and human resources into its investigation. On the other hand, if we, as a field, need to jettison a cherished scientific idea, let us do it after restrained analysis, lest we make further errors.

The article is structured as follows: First, we give a brief introduction of ego depletion. Second, we discuss the recent major threats to this idea. We then consider if, at least in principle, it is possible that the true ego depletion effect may be indistinguishable from zero after all—despite much seemingly supportive evidence in the literature. In the main fourth part of the article, we discuss the merits and faults of six arguments that could be brought forward in defense of ego depletion with respect to the previously discussed threats and beyond. The benchmark for these arguments was to determine (a) whether a proponent of ego depletion could be convinced by the earlier discussed threats that the phenomenon *does not* exist, and (b) whether a researcher skeptical of ego depletion research could be persuaded that, in fact, ego depletion *does* exist. In the general discussion, we outline the necessary next steps to reduce uncertainty and to provide more conclusive answers to currently pressing questions than is possible based on the extant evidence. We also examine to what extent publication bias and *p*-hacking may have contributed to bias in the ego depletion literature and the implications this may have for other research fields.[1]

What is Ego Depletion?

After narratively reviewing the literature on self-regulatory failure in diverse situations, behavioral domains, and populations, Baumeister and colleagues (Baumeister & Heatherton, 1996; Baumeister, Heatherton, & Tice, 1994) reached two conclusions: First, self-control is a domain-general construct. People who failed in self-control in one domain often appeared to do so in other domains as well. Second, exerting self-control has a psychic cost that increases chances of self-control failure in further attempts. Baumeister and colleagues observed that self-control failure often occurred after experiencing stress, or at night (after a demanding and tiring day) rather than in the morning. To formally test this hypothesis in controlled laboratory experiments, Baumeister and colleagues had participants perform two sequential self-control tasks and measured their performance on the second task as a function of whether or not the first task was effortful and demanding (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Muraven, Tice, & Baumeister, 1998). These and similar studies developed into a large and prominent literature in (social) psychology, the literature on ego depletion.

Ego depletion refers to the phenomenon that people perform poorer on a self-control task after having already engaged in a previous task requiring self-control. Thus, the hypothesized cause-effect relation is that the exertion of self-control in a first task (i.e., control of dominant responses such as thoughts, emotions, behavioral impulses and habits; the *cause*) leads to impaired subsequent self-control performance (the *effect*) compared to a control group that did not exert as much self-control in the first task. For example, people who had suppressed their emotions or thoughts in an initial task, resisted tempting food or controlled their attention, subsequently showed less persistence, ate more palatable (but unhealthy) food, drank more alcohol, or were more likely to smoke (for reviews and meta-analyses, see Carter, Kofler, Forster, & McCullough, 2015; Hagger, Wood, Stiff, & Chatzisarantis, 2010; Hirt, Clarkson, & Jia, 2016).

It is important to realize that in a multi-causal world self-control exertion is not necessarily the only aspect

[1] For full disclosure: All of us have done and are currently conducting ego depletion research. Some of us believe in the ego depletion effect but are uncertain about underlying mechanisms. Some believe in the general idea of demanding activities leading to poorer subsequent self-control performance, but are skeptical of laboratory ego depletion effects. Some no longer know what to believe in.

of a given manipulation that leads to impaired self-control performance. It is impossible to rule out or control for all other aspects that also exert causal influences on the hypothesized effect (Fiedler, 2017). Indeed, this notion has long entered theorizing about ego depletion: While it was initially assumed that only the exertion of self-control will cause a decline in subsequent self-control performance, the idea was broadened over the years. Now it is known that other factors may cause ego depletion effects that do not strictly conform to the narrow definition of self-control, for example engaging in any mentally demanding or effortful task (e.g., making decisions, solving complex arithmetic problems) and the engagement of executive functions more generally (Baumeister & Vohs, 2016b). Of course, not all other causal factors necessarily work in the same direction. That is, a given manipulation may not only evoke processes leading toward depletion effects, but may inadvertently also trigger processes that work against depletion effects (e.g., it may increase interest and motivation for the study, activate a mastery motive in participants). Thus, some processes may work against the effect and thus mask a true cause-effect relationship (Fiedler, 2017).

Ego depletion originally came out of social psychology, but was quickly applied to and investigated in various fields of psychological science, such as personality (Baumeister, Gailliot, De Wall, & Oaten, 2006), consumer behavior (Friese, Hofmann, & Wänke, 2008; Vohs & Faber, 2007), decision making (Pocheptsova, Amir, Dhar, & Baumeister, 2009), neuroscience (Heatherton & Wagner, 2011; Inzlicht & Gutsell, 2007; Luethi et al., 2016), cognitive psychology (Healey, Hasher, & Danilova, 2011; Robinson, Schmeichel, & Inzlicht, 2010), or organizational behavior (Christian & Ellis, 2011). What is more, ego depletion became one of psychology's export hits to the general public: The effect and its presumed implications for everyday life were prominently featured in countless public media reports and popular science books (e.g., Baumeister & Tierney, 2011; McGonigal, 2012).

Note that the definition of ego depletion introduced above refers exclusively to the behavioral phenomenon of impaired self-control performance after the previous exertion of self-control (i.e., functional level of analysis, De Houwer, 2011). This definition is agnostic about the psychological mechanism(s) underlying this behavioral effect (i.e., cognitive level of analysis, De Houwer, 2011). This is especially important to keep in mind as the term ego depletion conflates this behavioral, functional level of analysis with the process- or cognitive level by implying the reduction of a mental resource as the cause of the behavioral effect. Thus, although the term 'ego depletion' is unfortunate, we stick to it due to its widespread use in the published literature.

The term "ego depletion" alludes to the most widely-known model seeking to explain ego depletion effects, the strength model of self-control (Baumeister & Vohs, 2016b; Baumeister, Vohs, & Tice, 2007). The model posits that self-control relies on a limited and domain-independent resource that is partly depleted by any act of self-control, leaving the person more likely to fail in further attempts at self-control. Thus, on the cognitive level of analysis, this model identifies the reduction of a mental or physical resource as the crucial process underlying behavioral ego depletion effects. Despite abundant research on the topic, the nature of the resource remains elusive. Nevertheless, the model has inspired additional new literatures, for example, on the possibility to improve self-control by repeatedly producing small ego depletion effects (Beames, Schofield, & Denson, 2018; Friese, Frankenbach, Job, & Loschelder, 2017). In the meantime, several alternative theoretical models have sought to explain the ego depletion effect with potentially more readily measurable constructs such as motivation (Inzlicht & Schmeichel, 2012; Kool & Botvinick, 2014; Kurzban, Duckworth, Kable, & Myers, 2013). The details of these models are beyond the scope of this article and not relevant for its central purpose. Here, we are not primarily concerned with the theoretical explanation of the ego depletion effect (cognitive level). We are concerned instead with the more basic question of whether behavioral ego depletion is a real phenomenon in the first place (functional level, De Houwer, 2011).

Empirical Threats to the Ego Depletion Hypothesis

There are at least two major threats to the validity of the ego depletion hypothesis: First, a re-analysis of an early meta-analysis and a more recent meta-analysis concluded that the effect might not be different from zero (Carter et al., 2015; Carter & McCullough, 2014). Second, a large-scale registered replication report of one specific ego depletion study delivered a null effect on average (Hagger et al., 2016).

A first meta-analysis of published ego depletion studies in 2010 indicated a healthy mean effect size of $d = 0.62$ (Hagger et al., 2010)—a medium-to-large effect size according to convention and considerably stronger than most effect sizes reported in social psychology (Richard, Bond, & Stokes-Zoota, 2003). A re-analysis of the data investigated the presence of small study effects such as publication bias (Carter & McCullough, 2014). Publication bias occurs if studies with statistically significant results are more likely to be published than studies with null results. Correction for publication bias with trim-and-fill (Duval & Tweedie, 2000), the Precision Effect Test (PET; Stanley & Doucouliagos, 2014) and the Precision Effect Estimate with Standard Error (PEESE; Stanley & Doucouliagos, 2014) led to bias-corrected mean effect size estimates ranging from $g = -0.10$ to $d = 0.50$ (see the section on meta-analysis for details of these bias-correction techniques). Thus, despite the apparently strong evidence in favor of ego depletion in the Hagger et al. (2010) meta-analysis of published ego depletion studies (which did not include unpublished studies), this re-analysis found strong evidence for publication bias in the same dataset, questioning whether the true ego depletion effect is distinguishable from zero after all.

A second meta-analysis used stricter inclusion criteria, also included unpublished studies, and again conducted thorough analyses to detect publication bias (Carter et al., 2015). The analysis revealed an uncorrected mean effect size of $g = 0.43$. Correction for publication bias with trim-and-fill, PET, and PEESE led to corrected mean effect size estimates ranging from $g = -0.27$ to $g = 0.24$ (both significantly different from zero, albeit in opposite directions). Carter et al. (2015) concluded: "We find very little evidence that the depletion effect is a real phenomenon" (p. 796).

More recently, a large-scale, high-powered registered replication report (RRR; Hagger et al., 2016) sought to replicate one specific ego depletion study (i.e., Sripada, Kessler, & Jonides, 2014). This particular study was deemed appropriate for the RRR because the depletion manipulation was similar to manipulations that have been used in other depletion studies and because the complete study was computer-administered, a feature that was expected to minimize variability across labs (Hagger et al., 2016). A large-scale RRR can provide a high-powered test of a central prediction of a theory, if it employs a prototypical operationalization of central variables. In the RRR conducted by Hagger et al. (2016), a computerized variant of the widely-used e-crossing task was used as a manipulation of ego depletion (Baumeister et al., 1998), and performance on the Multi-Source Interference Task served as the dependent variable (MSIT; Bush, Shin, Holmes, Rosen, & Vogt, 2003). The protocol for this study was discussed and agreed upon with both the original authors and Roy Baumeister. Across 23 laboratories from different countries, this RRR revealed an overall null effect of the ego depletion manipulation on MSIT performance. This result constituted another significant threat to ego depletion research. Some recently published ego depletion studies, some of which were pre-registered, also found null effects (e.g., Etherton et al., in press; Lurquin et al., 2016; Osgood, 2017; Xu et al., 2014). Together, these findings led some researchers to conclude that ego depletion is an illusory phenomenon.

Is it Possible That Ego Depletion is Only an Illusion?

In contrast to the recent threats, hundreds of studies have attested to the existence of the ego depletion effect. The published literature is unusually large given the simple hypothesis: after a brief manipulation of group A exerting more self-control than group B, group A will perform worse than group B in a second task requiring self-control. Can such an extensive literature really be talking about a factually nonexistent phenomenon? How could that possibly be true? It could be true for at least two reasons: publication bias and *p*-hacking.

First, publication bias refers to the observation that studies with certain characteristics (e.g., significant results, large effect sizes) are systematically more likely to be published than studies without these characteristics. Publication bias may occur because researchers refrain from writing up and submitting studies that failed to produce the expected effect. In addition, if submitted, journals are less likely to publish non-significant work, because it does not conceptually replicate previous published studies (Bakker, van Dijk, & Wicherts, 2012).

Publication bias leads to an overestimation of population effect sizes. It is a crucial issue in the social sciences including psychology (Fanelli, 2010; Franco, Malhotra, & Simonovits, 2014, 2016). Ego depletion is no exception—there is evidence that the ego depletion literature suffers from strong publication bias (Carter et al., 2015; Carter & McCullough, 2014). For many years, large parts of psychological science have worked this way: Only 'clean', significant results were published. Reasons for failed (direct or conceptual) replications were primarily seen as indicating that replicating researchers had done something wrong. The possibility of false positive results in the extant literature was rarely considered.

Second, *p*-hacking refers to researchers engaging in questionable research practices to make their analyses statistically significant. It refers to making non-principled decisions to make one's data appear more robust than they actually are, where more principled and predicted analysis plans would have produced diverging, non-significant results (Simmons, Nelson, & Simonsohn, 2011). Common *p*-hacks are: reporting only dependent variables that 'worked' and omitting others, deciding for the inclusion or exclusion of outlying values dependent on which analysis reveals the 'better' results, peeking at the data during data collection and stopping when the desired pattern of results emerges without controlling for the increased Type-I error rate, or including covariates in an analysis without a clear theoretical rationale.

The prevalence of some practices encompassed by the term *p*-hacking is difficult to reliably determine, and it may vary by research field and culture, among many other variables. The prevalence of some *p*-hacks has likely been initially overstated (i.e., L. K. John, Loewenstein, & Prelec, 2012), with more appropriate measures suggesting much lower prevalence rates (Fiedler & Schwarz, 2016). Critically, though, even these lower prevalence rates are clearly higher than zero. *P*-hacking may become more likely when researchers feel personally tied to a hypothesis (Schaller, 2016), but it does not inevitably require bad intent or awareness of the consequences (Nelson, Simmons, & Simonsohn, 2018). In fact, most researchers likely underestimate how dramatically *p*-hacking can increase the likelihood of false-positives (Simmons et al., 2011). *P*-hacking leads to the overestimation of population effect sizes when the evidence is, in fact, weaker or possibly even nonexistent. Some *p*-hacking techniques are particularly likely to appreciably change results in small samples, which are typical for social psychology in general and ego depletion research as well.[2]

Taken together, with the knowledge we have today, there is little doubt that publication bias is present and that *p*-hacking has occurred in the ego depletion literature (as in many other literatures). Estimates of population effect sizes that do not take these factors into account will deliver overestimations. So, yes, Hagger and colleagues' (2010) estimate of $d = 0.62$ for the ego depletion effect is almost certainly an overestimate; but by how much? Is the correction required so large as to make the effect disappear altogether?

Arguments in Defense of Ego Depletion: Merits and Faults

The previous sections made clear that the current state of ego depletion research is messy; some results seem hard to reconcile with a robust effect. But is the empirical foundation for a critical stance substantial enough to convince a proponent that ego depletion is not real? Which arguments, if any, can be brought forward in defense of ego depletion? More, would these arguments convince even a skeptic? We consider six arguments and discuss their merits and faults in providing conclusive evidence for ego depletion: 1) limitations of meta-analyses and the RRR; 2) shortcomings of ego depletion manipulations and dependent variables; 3) moderator and mediator studies; 4) the absence of reverse depletion effects; 5) the size of the hypothetical file drawer; and 6) evidence for ego depletion in everyday life.

1) Limitations of meta-analyses and the RRR

Meta-analyses. The meta-analysis by Carter and colleagues (2015) presents a significant threat to the field of ego depletion research. It suggests that the effect may not be different from zero. Carter et al. (2015) found 620 ego depletion studies. In their analysis, the authors focused on the most frequently used independent and dependent variables. Studies not featuring one of these independent and dependent variables were not included. This led to 116 studies in the final data set (< 20% of the total). Carter and colleagues reasoned that these studies provided the most direct test of the ego depletion hypothesis. While this may be true, the restriction limited the possibility of investigating potential moderator effects that could have elucidated important theoretical, methodological, or procedural boundary conditions for the presence and magnitude of the ego depletion effect. Thus, the Carter et al. (2015) meta-analysis incorporated some of the most crucial studies to evaluate the ego depletion hypothesis, but it is limited in making broad conclusions about the ego depletion literature as a whole.

A second criticism that has been brought forward against the meta-analysis is the inclusion of many unpublished studies from relatively few different researchers (Cunningham & Baumeister, 2016). About one fourth (28%) of all included studies was conducted by just ten graduate students with little research experience. Unpublished theses and dissertations had smaller sample sizes than published studies and therefore had lower statistical power, which may partly explain why unpublished studies produced statistically significant effects less often. In short, proponents of ego depletion expressed strong concern that a small group of inexperienced researchers may have produced low quality studies that may have unduly biased the meta-analytic effect size estimate downwards (Cunningham & Baumeister, 2016).

We resonate with the notion that experience is important in designing and conducting high-quality studies. It would make little sense to assume that experience did not make us better at what we do. It is important to bear in mind, however, that graduate students are typically advised by more experienced researchers. In fact, graduate (and undergraduate) students in our own labs run many of our studies. We trust that by close supervision at all stages of the research process these relatively unexperienced researchers will produce high-quality studies. Thus, expertise may matter—particularly if supervision is superficial or incompetent—but various factors may have contributed to the smaller effect sizes of unpublished studies aside from expertise. Whether unpublished or published studies are less biased or biased by different factors is impossible to determine.

The meta-analysis spells out and provides evidence for what seems obvious today: There is publication bias in the ego depletion literature. It is impossible to say how severe this publication bias is and whether the true effect size based on all known and unknown studies (i.e., the file drawer) is indeed zero. Several techniques aim to correct for publication bias and provide bias-corrected effect size estimates (Duval & Tweedie, 2000; Hedges, 1984; Iyengar & Greenhouse, 1988; Simonsohn, Nelson, & Simmons, 2014; Stanley & Doucouliagos, 2014; van Assen, van Aert, & Wicherts, 2015).

To date, three of these procedures have been applied to the ego depletion literature (*p*-uniform, van Assen, van Aert, & Wicherts, 2015; PET and PEESE, Stanley & Doucouliagos, 2014; and trim-and-fill, Duval & Tweedie, 2000). Trim-and-fill is an iterative algorithm based on funnel plot asymmetry (a scatterplot with effect size on the *x*-axis and standard errors of the effect size on the *y*-axis). Asymmetrical funnel plots are assumed to indicate publication bias, because small, non-significant effect sizes and effect sizes in the 'wrong' direction are more likely to be excluded from the published literature. The trim-and-fill algorithm removes effect sizes until symmetry is restored and then imputes mirror images of the removed effect sizes. The summary effect based on all effects including the

[2] For a vivid demonstration of the astounding 'efficiency' of *p*-hacking, see http://shinyapps.org/apps/p-hacker/ (Schönbrodt, 2015).

imputed ones is then taken as an estimate of the unbiased true effect. The trim-and-fill procedure estimated the overall effect size at $d = 0.50$, $CI_{95}[.44, .56]$ for the Hagger et al. (2010) dataset (Carter & McCullough, 2014) and at $g = 0.24$, $CI_{95}[.13, .34]$ in the Carter et al. (2015) dataset—by Cohen's conventions moderate and small effect sizes, respectively.

Trim-and-fill makes several problematic assumptions: (a) publication bias is driven by weak effects rather than statistical significance (i.e., the assumption that weak effects are less likely to be published independent from statistical significance while in reality, even weak effects are published as long as they are statistically significant and even large effects are difficult to publish if not significant), (b) effect sizes are homogenous and distribute around one true fixed effect, and (c) asymmetry is necessarily caused by publication bias and not by other, potentially mundane small-study effects that cause true heterogeneity in effect sizes (Sterne et al., 2011). These assumptions are often violated in psychological research, which can lead to heavy under- and overestimations of publication bias. Several researchers argue that trim-and-fill should therefore not be used anymore (Carter, Schönbrodt, Gervais, & Hilgard, 2017; Moreno et al., 2009; Terrin, Schmid, Lau, & Olkin, 2003).

In the Precision-Effect Test (PET) and Precision-Effect Estimate with Standard Error (PEESE) effect sizes are regressed on a measure of precision (i.e., standard error of effect size in case of PET and square-root of the standard error in case of PEESE) in a weighted least-squares regression. The predicted estimate at perfect precision (e.g., standard error = 0) is taken as an estimate of the unbiased total effect. In the Hagger et al. (2010) meta-analytic dataset, the PET estimate was non-significant ($d = -0.10$, $CI_{95}[-.23, .02]$) and the PEESE estimate was positive and significant ($d = 0.25$, $CI_{95}[.18, .32]$, see Carter & McCullough, 2014). In the Carter et al. (2015) dataset, PET estimated a significant negative effect of ego-depletion (i.e., *improved* self-control after an initial exertion, $g = -0.27$, $CI_{95}[-.52, -.01]$), whereas the PEESE estimate was non-significant ($g = 0.003$, $CI_{95}[-.14, .15]$).

PET and PEESE are problematic because their effect size estimates vary widely under conditions typical for psychological research (i.e., heterogeneity, *p*-hacking, and publication bias), leading to both severe under-corrections and over-corrections (Carter et al., 2017). Some researchers therefore recommend avoiding PET and PEESE altogether (Carter et al., 2017).

Finally, the *p*-value-based selection method *p*-uniform (van Assen et al., 2015) assumes that (a) effect sizes are homogeneous, (b) non-statistically significant studies are uninformative, and (c) significant studies have the same likelihood of publication regardless of the exact *p*-value. *P*-uniform has recently been applied to the Hagger et al. (2010) and Carter et al. (2015) meta-analytic datasets (Blázquez, Botella, & Suero, 2017). The estimated effect sizes were $d = 0.64$ and $d = 0.66$, respectively. However, the method performs poorly and exhibits an upward bias in effect size estimates when the aforementioned assumptions (a) – (c) are violated (Carter et al., 2017; McShane, Böckenholt, & Hansen, 2016). In addition, the re-analyses by Blázquez and colleagues (2017) may have suffered from an unknown degree of bias, because some studies were included based on significant interactions (along with a significant depletion simple effect). In case of ordinal (attenuated) interactions, barely significant interactions may have come along with highly significant simple effects of ego depletion, which may have biased the effect size estimates upward. Another re-analysis using a similar technique came to lower estimates and different conclusions (Yost, 2016). Carter and colleagues (2017) recommend refraining from employing *p*-uniform.

In sum, the bias correction techniques that have been applied to ego depletion research suffer from a number of limitations. They are likely biased under many circumstances and often do not converge (Carter et al., 2017; Inzlicht, Gervais, & Berkman, 2015). This observation obviously does not provide evidence in favor of ego depletion either. Rather, it suggests that researchers should refrain from drawing far-reaching conclusions about the (non-)existence of ego depletion based on these methods. That being said, the presence of publication bias in the ego depletion literature appears beyond doubt to us. Fortunately, the development and validation of bias correction techniques is an active field of research, which may allow for more confident conclusions about the consequences of publication bias on effect size estimates as refined correction techniques become available.

RRR. With respect to the RRR (Hagger et al., 2016), critics argue that the study protocol did not use prototypical operationalizations of central variables. Specifically, the ego depletion manipulation (a variant of the widely-used e-crossing task; Baumeister et al., 1998) lacked a habit-creating first phase that may be important to increase self-control demands in a second phase (Baumeister & Vohs, 2016a). Indeed, the e-crossing task without habit-creating phase has been rarely used in previous research, although paper-pencil variants have produced expected depletion effects on self-control performance in a few studies (Tyler & Burns, 2009; Wan & Sternthal, 2008). Another potentially important deviation from previous research was that the e-crossing task was computerized while most applications of this manipulation have relied on the paper-pencil version. That said, recent evidence suggests that the e-crossing task can indeed lead to decrements in performance over time (within the same task), even without a prior habit-formation phase and even if the task is computer-administered, at least in variants somewhat longer than the one used in the RRR (Arber et al., 2017).

An analysis of the manipulation check data in the RRR showed that participants in the depletion condition did not report to be more tired after the initial self-control task than control participants (Dang, 2016). Note that the ego depletion hypothesis suggests that the exertion of self-control (i.e., control of dominant responses) causes subsequent decrements in self-control performance (functional level, De Houwer, 2011). It does not require the conscious experience of fatigue as a mediating mechanism (cognitive level). Nevertheless, based on the finding that the participants in the depletion group were no more fatigued than participants in the control group some researchers questioned whether the manipulation was strong enough (e.g., Baumeister & Vohs, 2016a). Of course, if the experimental manipulation was unable to induce the desired state, the experiment cannot actually test the central hypothesis.[3] Note, however, that the manipulation checks for experienced effort, difficulty and frustration were successful and showed large effect sizes, suggesting a successful ego depletion manipulation. That is, three out of four manipulation check items indicated that the manipulation was working as intended. Please note that this conclusion rests on the assumption that these four items in fact are indicative of the true cause that triggers ego depletion effects – an assumption that conflates the functional with the cognitive level of analysis (De Houwer, 2011, see also the final section for a discussion of appropriate manipulation checks in ego depletion research).

Supplementary analyses indicated that participants in the depletion condition who invested relatively more effort (+1 *SD* above the mean) performed more poorly on the dependent variable than participants who invested relatively little effort (Dang, 2016). However, this association was no longer present when participants from both conditions were examined (after all, some control participants likely felt they invested effort as well). In addition, none of the other three manipulation check items (difficulty, fatigue, frustration) showed a similar pattern. We are thus wary of drawing strong conclusions from this analysis.

For the second self-control task (the dependent variable), the RRR employed the MSIT (Bush et al., 2003). The MSIT, or variants thereof, has rarely been used as a dependent variable in ego depletion research (Shamosh & Gray, 2007; Wenzel, Kubiak, & Conner, 2014), which was later criticized (Baumeister & Vohs, 2016a). However, the task bears similarity with the widely-used Stroop task (Stroop, 1935). It was therefore plausible to expect performance impairments after ego depletion.

Despite the capability of an RRR to test a central prediction of a theory with great statistical power across several laboratories, the present format was restricted to one specific combination of ego depletion manipulation and dependent variable. The RRR strongly suggests that the present combination of digitalized e-crossing and MSIT does not produce a reliable ego-depletion effect. This null effect constitutes a significant threat to the ego depletion hypothesis. Note, however, that one salient characteristic of ego depletion research is the combination of various IVs with various DVs across studies. The ability of one particular IV-DV combination to validate or 'disprove' the general idea of ego depletion is necessarily limited. It appears premature and logically unjustified to dismiss the phenomenon and the complete existing ego depletion literature based on a single IV-DV combination in one specific study, even a statistically powerful one.

Interim conclusion. Current meta-analytic techniques to correct for publication bias suggest vastly different bias-corrected average effect sizes ranging from $g = -0.27$ to $d = 0.66$. These techniques do not converge and are therefore unable to provide reliable bias-corrected effect size estimates. The effect could be zero, but it could reasonably also be larger than zero. With respect to the RRR, there is some doubt whether the experimental manipulation was strong enough. Nevertheless, it seems premature to dismiss the entire ego depletion hypothesis based on a single failed IV-DV combination. In all, proponents of ego depletion are unlikely to change their conviction as a result of criticisms that are based on the discussed bias-correction techniques and the single IV-DV combination from the RRR.

2) Shortcomings of ego depletion manipulations and dependent variables

One truism of experimental work is that researchers can only expect reliable effects of their experimental manipulation on a dependent variable if this manipulation successfully manipulated the construct of interest. Proponents of ego depletion research might argue that many studies that did not show a depletion effect failed to meet this requirement (e.g., Baumeister & Vohs, 2016a). In the meta-analysis of 198 ego depletion studies by Hagger and colleagues (2010), only 60 (30.3%) included a manipulation check asking for difficulty of the first task, only 31 (15.7%) asked for subjective effort, and only 25 (12.6%) for fatigue after the first task. Remember that according to the definition, the exertion of self-control is the causal factor that presumably leads to decrements in subsequent performance. However, when scientists do not know whether and to what extent participants exerted control, the success of the experimental manipulation is in doubt.

In addition, the variety of ego depletion manipulations is great. They vary from very brief manipulations such as 20 incongruent (depletion condition) versus 20 congruent (control condition) Stroop trials (Yam, Chen, & Reynolds, 2014), to

[3] There actually are occasional studies that produced a significant depletion effect without a significant manipulation check (e.g., Finkel et al., 2006; Pocheptsova et al., 2009; Segerstrom & Nes, 2007). This may indicate that ego depletion effects are not necessarily preceded by conscious perceptions of having invested more effort, or conscious feelings of mental fatigue, etc. Alternatively, one may argue that some depletion effects may not be due to participants having exerted effort but to unintended demand-, expectation-, experimenter-, or other unknown effects. In a multi-causal world, more than one cause can lead to behavioral patterns resembling ego depletion effects.

completing several demanding tasks in a row in the depletion condition, each lasting several minutes (Vohs, Baumeister, & Schmeichel, 2013). Some manipulations may—if at all—only be capable of inducing spurious motivational shifts. In light of many everyday chores that are much more effortful than some of the very brief manipulations found in the literature, it is difficult to discern how these manipulations could exert reliable effects on subsequent self-control performance. In contrast, more elaborate manipulations require more profound exertion of control (possibly leading to consciously-experienced fatigue) and may thus more likely impair subsequent performance.

In a related vein, it is unclear how suitably the large variety of dependent variables can detect ego depletion effects. The outcomes that have been employed vary from Stroop tasks that require the control of dominant responses and are therefore prototypical for prominent conceptualizations of self-control (Baumeister, 2014) to persistence on unsolvable puzzles (Baumeister et al., 1998) and anagrams (Dvorak & Simons, 2009), to handgrip performance (Muraven et al., 1998) and consumption of tempting but unhealthy food (Friese et al., 2008) or alcohol (Christiansen, Cole, & Field, 2012; Muraven, Collins, & Nienhaus, 2002), to smoking (Heckman et al., 2017; Shmueli & Prochaska, 2009), to arithmetic calculations (Vohs et al., 2008) and changes in choice behavior (Pocheptsova et al., 2009), to aggressive intentions and behavior (DeWall, Baumeister, Stillman, & Gailliot, 2007; Stucke & Baumeister, 2006), to risk-taking behavior (Freeman & Muraven, 2010), and to achievement motivation (Imhoff, Schmidt, & Gerstenberg, 2014). While these examples are still but a subset of the universe of employed outcome variables in the depletion literature, it is already apparent that outcome variables vary widely and some do not seem to have much in common conceptually. Are they all really measuring the same thing—self-control?

Meta-analytic findings suggest that a diverse array of variables presumably indicating self-control share a common core, but a rather small one (Duckworth & Kern, 2011). Each outcome variable assesses additional aspects besides self-control. Had independent and dependent variables been randomly allocated to studies, depletion effects in many of those studies would speak to the generalizability of the hypothesis and suggest that it may really be the common core of the variables—self-control—that makes these variables susceptible to influences of self-control exertion during the first task. However, independent and dependent variables were not randomly allocated, but may have been selected strategically in a way researchers felt they would ‘work’ (Bless & Burger, 2016). Despite these cautious comments, it seems fair to say that many common outcome variables tap into some form of self-control in the sense of an effortful regulation of thoughts, emotions, impulses, habits, or behaviors.

In addition to validity concerns, the reliability of dependent variables is often unknown and hardly ever reported in ego depletion research. It is possible that low reliability of dependent variables and high measurement error may have methodologically reduced the detected effect sizes in many studies, thus underestimating a true population effect and effectively leading to false negative findings.

Interim conclusion. Many ego depletion studies may have offered no real test of the postulated effect—the manipulations were too weak to reasonably assume reliable effects. Manipulation checks have been rarely employed. In addition, the validity and reliability of most dependent variables is unknown and may have compromised effects.

Evidently, the prior observations do not provide evidence in favor of ego depletion. If anything, they undermine trust in the ego depletion literature. However, weak manipulations and/or unreliable/invalid dependent variables may explain failures to replicate ego depletion effects. The data of those studies necessarily remain inconclusive; they are neither able to provide evidence for the existence nor the non-existence of ego depletion. Stronger manipulations (e.g., Vohs et al., 2013) that reliably produce pronounced effects on manipulation checks are required in future studies to allow for solid tests of the hypothesis. Future meta-analyses may want to run sub-analyses with studies that produced strong effects on manipulation checks. They should be more likely to produce reliable and larger effect sizes than studies without supportive manipulation checks.

3) Moderator and mediator studies

A typical ego depletion study features a one-factorial between-participants design with two experimental conditions (control versus depletion). In this design, ‘genuine’ false positives that support the theory occur in 2.5% of all studies if the true effect is zero. In addition, *p*-hacking the data in the preferred direction is relatively easy, especially if the sample size is small (which is true for many published ego depletion studies). Thus, effectively, *p*-hacking increases the rate of false positives to an unknown extent beyond the rate of 2.5% genuine false positives in these studies.

Experimental designs in moderator and mediator studies are more complex. Genuine false positives (i.e., false positives not produced through *p*-hacking) are less likely in these cases because a specific pattern in a more complex study design is less likely to occur by chance than a mean difference between two conditions (Murayama, Pekrun, & Fiedler, 2014). Does a more complex study design also imply that it is more difficult to *p*-hack the results of these studies than of two-condition studies? To the extent this was true, moderator and mediator studies may provide more robust evidence for ego depletion than two-condition studies.

Moderation. More than one hundred moderator studies have investigated factors that either counteracted an induced ego depletion or prevented it from occurring in the first place (for an overview, see Loschelder & Friese, 2016). These moderators are either experimentally manipulated or individual difference variables. For example, incentives to perform well in the second task counteract ego depletion effects (Muraven & Slessareva, 2003). Incentives can either be financial or social (believing that one's efforts will benefit others). Similarly, participants who exerted self-control and were then led to be in a positive mood by watching a funny video or receiving a surprise gift did not show a depletion effect (but people put in neutral or sad mood did; Tice, Baumeister, Shmueli, & Muraven, 2007). Other examples of factors counteracting ego depletion effects are a high construal level (Agrawal & Wan, 2009; Schmeichel & Vohs, 2009) and implementation intentions (Webb & Sheeran, 2003). A prominent example of a moderator presumably preventing depletion to unfold are lay theories about willpower. This approach assumes that people tend to behave in line with their personal beliefs. People who dispositionally believe or are experimentally led to believe that their willpower is limited show the ego depletion effect. However, if they believe or are led to believe that willpower is unlimited, they do not show the depletion effect (Job, Dweck, & Walton, 2010). Similarly, action (vs. state) oriented individuals do not show depletion effects, presumably because they continue allocating resources to difficult tasks when state-oriented individuals conserve strength such as after a demanding initial task (Gröpel, Baumeister, & Beckmann, 2014). Trait self-control has also been found to moderate the ego depletion effect in several studies, but the jury is still out on the direction of this moderator effect: In some studies, high trait self-control was associated with more (Imhoff et al., 2014), in other studies it was associated with less susceptibility to ego depletion effects (DeWall et al., 2007).

In many studies, experimentally manipulated moderators are investigated in 2 (ego depletion: yes vs. no) × 2 (moderator: low vs. high) experimental designs or one-factorial designs with three conditions (control, depletion, depletion+moderator). Assuming that ego depletion does not exist and acknowledging that moderators are often unable to improve self-control performance beyond baseline levels (Schmeichel & Vohs, 2009) leads to the expectation of no differences across conditions, at least not beyond random sampling error. A researcher aiming to find evidence for a moderator would have to *p*-hack the mean in one cell only—the one depletion condition expected to show impaired performance. Presuming this was the malicious or subconscious intent of a researcher, it would be relatively easy to *p*-hack such a result pattern.

In studies investigating individual difference moderator variables, the data are typically analyzed with multiple regression. An interaction is expected to show a depletion effect for people with low, but not high values of the moderator variable (or vice versa). Put differently, a correlation between the individual difference variable and the outcome is expected in the depletion condition, but not the control condition. We are not aware of any simulation work suggesting that *p*-hacking a non-existent correlation is more difficult than *p*-hacking a non-existent mean difference. Hence, *p*-hacking individual difference moderator studies may be similarly easy as *p*-hacking experimentally manipulated moderator studies.

Mediation. Mediator studies are less prevalent than moderator studies. In one study using electroencephalography (EEG), emotion suppression led to poorer subsequent performance in a Stroop task (Inzlicht & Gutsell, 2007). This ego depletion effect was partially mediated by a reduced error-related negativity during the Stroop task, a neural indicator of conflict and error monitoring. The implication of this study was that exerting effort may impair monitoring processes, which may at least partially account for impaired subsequent performance (see also Dang, Bjorklund, & Backstrom, 2017). One series of studies found evidence for the idea that reductions in self-efficacy with respect to the upcoming task would mediate ego depletion effects. This effect was further moderated by lay theories of willpower: Reduced self-efficacy was particularly observed in individuals who believed that willpower is limited (Chow, Hui, & Lau, 2015; see also Fischer, Greitemeyer, & Frey, 2007; Graham, Ginis, & Bray, 2017). Other studies found stronger ego depletion effects, the more subjective exhaustion participants reported after the first task (e.g., Govorun & Payne, 2006; Job et al., 2010; Yam et al., 2014).

Mediator studies are more complex to *p*-hack than moderator studies because they make more specific assumptions about the data pattern. The experimental condition (control vs. depletion) is expected to correlate with both the dependent variable and the mediator, which in turn is correlated with the dependent variable. Mediation occurs if the indirect effect of the ego depletion manipulation (partially) explains the direct effect of the experimental condition on performance (Hayes, 2013). That is, when accounting for the effect of the mediator on the dependent variable, the correlation between the experimental condition and the dependent variable becomes significantly smaller (Baron & Kenny, 1986; Spencer, Zanna, & Fong, 2005). While it is likely possible to *p*-hack a statistical trend for a mediation toward significance, everything else being equal, the necessary overall correlational pattern—including shared variance of mediator and IV/DV—is more difficult to make up from scratch than *p*-hacking a significant result in two-condition or moderator studies. The required set of assumptions for statistical mediation patterns is more complex than what is required for basic two-condition or moderator studies.

Interim conclusion. Due to the more complex study designs, genuine false positives are less likely to explain the occurrence of moderators than they are to explain differences in two-condition studies. Nevertheless, the high prevalence of moderators does not constitute a viable defense of ego depletion, because moderation studies are similarly susceptible to *p*-hacking than two-condition studies assuming a true effect of zero. Studies showing statistical mediation may be interpreted as indirect evidence for a true ego depletion effect larger than zero because such data patterns are relatively more complex to produce through *p*-hacking alone if the necessary correlations are not already present in the data set.

These tentative conclusions in defense of ego depletion come with three caveats. First, it is possible that mediator effects are less stable than the literature suggests. Hardly any published depletion studies have been pre-registered. It is thus unknown in how many studies potential mediators (and also moderators) were assessed but not reported (as was common practice not only in ego depletion research for decades). Considering the size of the ego depletion literature, it is startling that only a few studies investigated underlying mechanisms and reported statistical mediation. Second, many moderator and mediator studies (including some of our own) have embarrassingly low sample size. In such small samples, seemingly minor variations in data analysis may have substantial effects on statistical significance. Thus, the increased difficulty to alter results through creative data analysis in mediation studies is at least partly offset by the ubiquity of small samples.

Third, small sample sizes are problematic for another reason. Although true effect sizes for given combinations of ego depletion manipulation and dependent variables are unknown, assuming an average small-to-medium effect size is prudent at this point in time. Small effects require large samples. To give an example, to detect an effect of $d = 0.35$ with at least 80% power ($\alpha = .05$, two-tailed *t*-test), 260 participants are needed. If a moderator is assumed to completely eliminate the effect, the required sample size for a 2×2 design is a whopping 1,040 (Simonsohn, 2014). Assuming a depletion effect of $d = 0.5$, the required sample size for moderation is still 512. The largest study reported in the Hagger et al. (2010) and Carter al. (2015) meta-analyses included 501 participants (Pocheptsova et al., 2009, Study 2), the second largest 284 participants (Pocheptsova et al., 2009, Study 1). Neither of these were moderator studies. This implies that it is unlikely that such a high number of studies would find significant moderation without bias, given what we know about the probable size of the depletion effect and statistical power to detect moderation of it.

In sum, we believe that, in principle, mediator studies can provide stronger evidence for ego depletion than studies making less specific assumptions about the data. The existing evidence is limited, however, and does not allow for far-reaching conclusions.

4) Absence of reverse depletion effects

If the true ego depletion effect was indeed zero, empirical studies should not consistently reveal null effects. Instead they should distribute symmetrically around the mean of zero with larger, statistically more powerful studies aligning more closely around zero due to smaller sampling error. In meta-analyses, this expectation is often inspected with funnel plots (Borenstein, Hedges, Higgins, & Rothstein, 2009; Sterne & Egger, 2001). Figure 1 shows two funnel plots (Panel A and B) of fictitious studies with an underlying true population effect of zero. For a true population effect of zero, these simulations show that, by chance, statistically significant studies will occur on either side of the 95% funnel. Hence, a proponent of ego depletion might argue that the lack of reverse depletion effects is indicative of the true population effect being larger than zero. Put differently, if publication bias were the main reason for why the published literature—comprising several hundred significant studies to the right of the funnel—reports almost exclusively positive depletion effects (i.e., poorer performance after exerting self-control), there should be hundreds of studies showing reverse depletion effects (i.e., *better* performance after exerting self-control; see left half of Figure 1A). In the current debate and given the extensive skepticism that the ego depletion literature has received, we would expect reverse depletion effects (especially significant ones) to be easily publishable. Several journals would publish such studies without hesitation as long as they are methodologically sound. However, such studies are hard to find (cf. Converse & DeShon, 2009; Tuk, Zhang, & Sweldens, 2015; Van Reet, 2015) and when they are found, they are sometimes not theoretically unexpected, but rather predicted reverse depletion effect (e.g., because lay theories about willpower differ from one culture to another, Savani & Job, 2017).

As you may have noticed, the above reasoning is incomplete. Under the assumption of a true null effect, not all significant effects in the expected direction are 'genuine' false positive results. The consequences of *p*-hacking need to be taken into account. *P*-hacking will shift positive, but non-significant effect sizes toward even more positive results so that they reach statistical significance. This is true especially for small studies in which some *p*-hacks more easily produce appreciable effects. In Figure 1A, studies 'in danger of being *p*-hacked' are simulated as white dots. These are studies that reveal a non-significant tendency in the expected direction.[4]

[4] In the simulation that produced Figure 1A, we classified a study as being 'in danger of being *p*-hacked' when the effect size was positive and the *p*-value fell above .050 and below a pre-defined cut-off. Upper cut-off values were linearly decreased from $p = .500$ for studies with the lowest precision (i.e., highest standard error) to $p = .200$ for studies with the greatest precision. (We deemed it unlikely that researchers would *p*-hack studies to significance if the original *p*-value without *p*-hacking was larger than .500. We trust there are limits to self-deception that researchers may rely on when justifying their *p*-hacks, and also limits to what the vast majority of researchers deems morally acceptable to publish). This scenario assumes that it is easier to achieve greater changes in the *p*-value by *p*-hacking in smaller studies compared to larger studies. The probability for any study 'in danger of being *p*-hacked' to actually being *p*-hacked was set to 50%. It appeared reasonable to define this *p*-hacking probability as independent of precision, because two presumably opposing mechanisms may be at work: As studies become smaller, certain *p*-

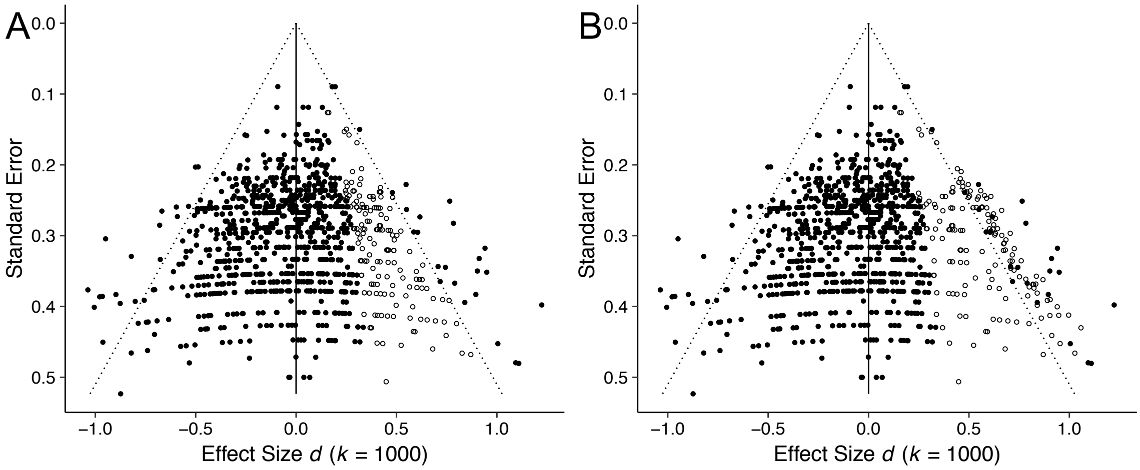

Figure 1. *Panel A*: Funnel plot of 1000 hypothetical studies with a true underlying effect size of zero. Studies with a smaller standard error (greater precision) cluster more closely around the true effect size of zero. Studies with a greater standard error cluster more widely. Studies within the funnel reveal non-significant results ($ps > .050$). Approximately 5% of all studies deliver either a significantly positive (~ 2.5%) or significantly negative (~ 2.5%) outcome (Type-I error; left and right area outside the funnel). White dots represent a subset of the 1000 total studies that are 'in danger of being *p*-hacked'. These studies revealed non-significant positive effects with *p*-values from close to significance ($p = .050$) up to $p = .500$ (see footnote 4 for details of the simulation). *Panel B*: Forty-six percent of the studies in danger of being *p*-hacked (white dots in Figure 1A) have now been *p*-hacked to *p*-values of .040 to .010. All studies 'in danger' had a 50% chance of being *p*-hacked. All other studies (typically the file drawer) remain unchanged. The black and white dots on the right of Figure 1B outside the funnel are likely to be published as supporting the hypothesis. For the ego depletion literature, assuming the effect does not exist, this group of studies that is significant by chance (i.e., Type-I error) or *p*-hacked comprises several hundred studies. See the main text for scenarios about the size of the hypothetical ego depletion file drawer.

The left half of the effect size distribution showing (at least descriptively) reverse depletion effects should be unaffected by *p*-hacking, because transforming a reverse depletion effect into a significantly positive effect seems practically unfeasible—a very severe amount of *p*-hacking would be necessary. Therefore, there will be fewer significant reverse depletion effects (i.e., only 'genuine' false positives on the left, no additional studies due to *p*-hacking) than there will be regular depletion effects (i.e., genuine false positives plus *p*-hacked effects outside the funnel on the right). Nevertheless, one reasonable scenario suggests that there should be around 190 significant reverse depletion effects (i.e., false positives) under the premise of a true null effect (see next section for details). Proponents and skeptics alike should wonder where all these reverse depletion studies are. One possible reason for the fact that only a few reverse depletion effects have been published to date could be that the true ego depletion population effect is larger than zero. True effects larger than zero strongly decrease the probability of significant reverse depletion effects. For example, the probability of a significant reverse-depletion effect is approximately 1.1% when the true (fixed) effect is $d = 0.1$ and approximately 0.5% when $d = 0.2$ instead of 2.5% when the true effect is $d = 0$ (based on the ego depletion literature sample size distribution, see footnote 4).

Interim conclusion. Given a true null effect, large numbers of reverse depletion effects should exist. The fact that only a small number of these have been published could be interpreted as indirect evidence that the true ego depletion effect may not be zero, but more likely distribute around a positive mean effect.[5]

hacks become increasingly effective in achieving changes in the *p*-values (i.e., increasing the probability of successful *p*-hacking), but as original *p*-values become larger, longer 'distances' to $p <= .05$ need to be covered (i.e., decreasing the probability of successful *p*-hacking). The concrete impacts of these two opposing tendencies is difficult to quantify. We therefore chose to assume a constant *p*-hacking probability across the various degrees of study precision. To simulate study sample sizes, we randomly drew k =1,000 values (with replacement) from the study sample sizes reported in Carter et al. (2015) and Hagger et al. (2010). This method thus approximates typical sample sizes found in ego depletion research. Total sample sizes ranged from $N_{\min} = 16$ to $N_{\max} = 501$ ($M = 55$, $Md = 47$, $SD = 41.67$). Studies that have been 'successfully' *p*-hacked in Panel B have been *p*-hacked to *p*-values of .040 to .010. Note that the various assumptions producing Figure 1 merely depict one of many reasonable scenarios for ego depletion research. Figure 1 is not intended to depict an exact estimation of the (unknown) reality. The R code can be found at https://osf.io/s79fe/.

[5] If ego depletion were a nonexistent phenomenon, there would be a fixed and single true depletion effect size of zero. In contrast, if ego depletion were a real phenomenon, talking about one single true effect size actually is misleading. The assumption of a single true effect size is unwarranted in most research literatures (Borenstein et al., 2009). Rather, there may be several true effect sizes depending on boundary conditions (random effects model). In the case of ego depletion research, such conditions are represented by, for example, the multitude of different depletion manipulations in terms of type, severity, and duration, different dependent variables, and their combinations. This observation also explains why some researchers regard the meta-analytic search for an estimation of 'the' mean depletion effect pointless – a notion that resonates with us to some extent. However, meta-analyses provide unique opportunities to reveal important moderators. Also, estimation of 'the' mean depletion effect becomes crucial, not pointless, if (valid) corrections for publication bias and *p*-hacking suggest that the mean effect may in fact be zero.

Alternatively, one would have to assume that studies showing reverse depletion effects do exist in large numbers, but are being kept from publication (by authors, reviewers, and/or journals). Hence, it would be valuable if researchers opened their file drawers of reverse (and non-significant) depletion effects to get a more accurate impression of the true evidence (e.g., by submitting them to journals or simply posting the data to public repositories).

5) Size of the hypothetical file drawer

The discussion on reverse depletion effects is also informative with regard to developing an idea about a plausible size of the hypothetical ego depletion file drawer. Under the assumption that the true ego depletion effect is zero, it would be impossible to tell which proportion of significant studies (outside of the funnel to the right in Figure 1B) is *p*-hacked (white dots) and which proportion represents non-*p*-hacked, 'genuine' false positives (black dots). Either way, the studies to the right, outside of the funnel are more likely published (as supportive evidence for ego depletion) than all other studies depicted in the funnel plot. Given that there are several hundred published studies (i.e., genuine false positives plus *p*-hacked studies, if the true effect were indistinguishable from zero), the figure illustrates that the file drawer of unpublished studies would have to be very large.

To quantify one hypothetical scenario: Let us assume there are 750 published ego depletion studies supporting the hypothesis (i.e., 750 black and white dots outside the funnel to the right in Figure 1B).[6] Let us further assume that all these significant studies are published, while ego depletion truly does not exist ($d = 0$). Accepting the assumptions that we entered in our simulation (see footnote 4), the results suggest that 74.8% of those 750 studies are *p*-hacked, and that 25.2% (= 189 studies) are genuine false positives. By definition, 'genuine' false positives on the right represent about 2.5% of all conducted studies when alpha is set to .05 in a two-tailed test. Hence, this scenario would imply a total number of 7,560 conducted studies (189 studies/2.5% × 100%). A rough estimate of the file drawer under these assumptions is 6,810 studies (7,650 total – 189 false positives – 561 *p*-hacked).

In the scenario depicted in Figure 1, studies 'in danger of being *p*-hacked' (i.e., white dots in Figure 1A) were hacked to significance with a probability of 50% (see footnote 4). If we assume that the probability of a study 'in danger of being *p*-hacked' will in fact be *p*-hacked is not 50% but 80%, the hypothetical size of the file drawer is reduced to 4,450 studies. If we set it a more conservative level, say 20%, the file-drawer increases to 12,970 studies.

Note that the intent of these hypothetical scenarios is not – akin to the much criticized fail-safe *N* approach (Rosenthal, 1979, see Lakens, Hilgard, & Staaks, 2016; Scargle, 2000) – to determine a number of null-findings necessary to bring down a hypothetical mean ego depletion effect to zero. Fail-safe *N* seeks to estimate a number studies with null effects that could be tolerated before researchers would have to conclude there is no effect. Thus, a critical assumption of this approach is that the significant studies that are already there provide valid evidence for the phenomenon of interest. Here, we take a different approach: We assume a null effect and treat the significant studies as false positives. In addition, we make specific assumptions about *p*-hacking. Based on these assumptions we develop an approximation of the size of the hypothetical file drawer based on simple arithmetic, given 2.5% genuine false positives in the expected direction and different degrees of severity of *p*-hacking.

Two further observations seem noteworthy: First, under the alternative premise that the ego depletion effect is real and thus truly larger than zero, the hypothetical file drawer would be of smaller size. In this scenario, the hypothetical size of the file drawer is inferred from the number of genuine (i.e. unhacked) positives and the mean power of the simulated studies (that we can infer from the sample size distribution, see footnote 4). Specifically, if we have an estimate for the number of effects that were significant before *p*-hacking ('true positives'), we can compute the number of studies that would need to be conducted to achieve that number of true positives at a given level of statistical power. For example, assuming a small true (fixed) effect of $d = 0.1$ reduces the file drawer to 3,510 studies (2,131 studies at $d = 0.2$; 1,295 studies at $d = 0.3$), when 50% of those studies 'in danger of being *p*-hacked' are indeed *p*-hacked. Put differently, if the file drawer was found to be considerably smaller than the numbers simulated above (between 4,450 – 12,970 studies), this might provide indirect evidence for a true population effect size greater than zero.

Second, up to this point, we assumed that all studies in support of ego depletion were published and that the file drawer exclusively consists of studies that do not support ego depletion. However, in fact, the file drawer may not only contain studies that failed to reveal significant ego depletion effects: One of the most frequent reasons for the rejection of manuscripts is that the reported findings are regarded not novel enough by journal editors and reviewers. Thus, it is rather likely that not all supportive evidence in favor of ego depletion

[6] In their literature search, Carter and colleagues (2015) identified a total of 620 studies through the end of 2012. We estimated that at least 400 of these were published. For the time span 2013-2017, we found 1600 citations of the following three major ego depletion references: Baumeister et al. (1998; 2007) and Muraven and Baumeister (2000). A research assistant went through the abstracts of all of these articles that were published in 2013, 2015, and 2017, counted the number of reported depletion studies, and extrapolated to the complete period of 2013-2017 (≈ 410 studies). Thus, a rough estimate of the total number of published ego depletion studies until the end of 2017 is 800. In the following calculations, we used the more conservative calculation of 750.

is published. Instead, an unknown number of supportive ego depletion studies is likely hidden in the file drawer. Essentially, if ego depletion was real and if these studies have not been *p*-hacked, they may constitute false negatives—missed opportunities to provide (published) evidence for a real effect.

This possibility can be incorporated in the file drawer estimations. Let us assume that only 70% of studies that reached significance and were submitted for publication, are eventually published. We again assume that 50% of studies 'in danger of being *p*-hacked' are *p*-hacked, and that there is no true ego depletion effect. In this case, the size of the hypothetical file-drawer of unpublished significant and non-significant studies would be increased to 10,050 studies, as opposed to 6,810 when all significant studies are published. Alternatively, we could assume that 50% or 90% of significant studies are published, which would yield a file drawer the size of 14,370 and 7,650, respectively.

Interim conclusion. Taken together, these simulations suggest that under the assumption that ego depletion does not exist, the file-drawer of unpublished studies would be very large. Determining the exact size of the file drawer is impossible. This depends on various factors such as the number of published studies, the severity of *p*-hacking, the true size of the effect, and the percentage of significant studies that are published: In our simulations, the theoretical file-drawer of studies not significant in the expected direction comprised between ~4,450 and ~12,970 studies (assuming that all significant studies in favor of depletion are published). Even the lowest estimate of the file drawer, assuming that 80% of all studies 'in danger of being *p*-hacked' end up being *p*-hacked, is quite substantial (~4,450 studies).

We hasten to say that these estimates strongly depend on the specific assumptions underlying the respective simulations (e.g., true effect of d = 0, no heterogeneity, see footnote 4). They also depend on further assumptions like the percentage of significant studies that are published and that all conducted studies are equally valid tests of the hypothesis. Our assumptions are necessarily false. They are an attempt to approximate the unknown reality; this attempt is open for refinement. However, we are confident that reasonable variations on the central parameters of the simulation would not lead to categorically different conclusions. So, the general aim of this section is not to provide realistic point-estimates for the size of the file drawer. This is impossible. The general aim is to point out that while skeptics of ego depletion may feel that a large number of file drawer studies is quite realistic, proponents of ego depletion may be unlikely to be convinced that publication bias and *p*-hacking could have been severe enough to produce the existing literature based on a true null effect.

6) Evidence for ego depletion in everyday life

Ego depletion research has overwhelmingly followed the sequential-task paradigm in controlled laboratory experiments. The general idea these studies aim to examine, however, is broader and certainly not bound to the laboratory. Remember that the initial impetus for ego depletion work was inspired by Baumeister and colleagues' (1994) observation that self-control failures often occur later during the day and under high regulatory demands, when people feel tired, fatigued, or otherwise deprived of resources. One possible defense of ego depletion is based on the grounds of everyday experience: Almost everyone has felt moments of fatigue and a resultant loss of self-control. These depletion experiences may be too tangible to be dismissed as an unreal phenomenon.

In the current discussion, we deem it helpful to distinguish between the basic theoretical idea of ego depletion ("After experiencing pronounced regulatory demands people show impaired self-control performance") and the laboratory research that has tested this idea ("Are there reliable ego depletion effects using the sequential-task paradigm?"). The general idea may still have merit even if its predominant operationalization – the laboratory sequential-task paradigm – may have failed to produce reliable and robust ego depletion effects.

Indeed, some real-world evidence supports the general idea that high demands impair subsequent behavior that requires people to overcome an easy, dominant response. For example, a large-scale longitudinal field study of over 4,000 caregivers working in hospitals revealed that hand hygiene compliance rates dropped by almost 9 percentage points from the beginning to the end of a typical work shift (Dai, Milkman, Hofmann, & Staats, 2015). Effects were stronger when work intensity was high and longer breaks between shifts buffered the effect. Another study investigated the prescription rate of antibiotics in response to acute respiratory infections (Linder et al., 2014). Prescribing an antibiotic was considered an easy and safe option for clinicians. In line with general notion of ego depletion, clinicians prescribed increasingly more antibiotics within the course of a morning shift up to lunch. The same pattern occurred after lunch towards the afternoon.

In the academic domain, a large-scale study in Denmark showed that public school students performed worse for every hour later a test was taken, but having a 20-30-minute break before the test improved performance (Sievertsen, Gino, & Piovesan, 2016). In the consumer domain, two field experiments in the sphere of product customization of cars investigated the effects of making a series of easy choices (i.e., few options) versus complicated choices (i.e., many options). Making choices has been hypothesized to lead to ego depletion effects, because it requires the individual to repeatedly consider and weigh information and to actively choose between several alternatives the most promising option (Vohs et al., 2008). In these studies, having made a series of complicated (as compared to easy) decisions led consumers to more heavily rely on default options offered by the producer in subsequent decisions and let them shy away from actively choosing other offered options (Levav, Heitmann, Herrmann, & Iyengar, 2010). Finally, Augenblick and Nicholson (2016) found similar effects in the political domain. These authors investigated nearly one million voting outcomes for federal, statewide, and local contests in San Diego. They found that placing an issue later on the ballot—in other words, after more preceding voting decisions had already been made—increased the relative frequency of abstentions, the number of votes for the status quo (i.e., the default option). It also increased the tendency to vote for the first candidate listed in multicandidate decisions. The authors interpreted these findings as evidence for the assumption that making many choices led voters to experience so-called decision fatigue, which increased their tendency to choose the relatively easier option in later voting decisions.

Other research using experience sampling methodology investigated the ego depletion hypothesis more directly. Several of these studies used daily physical activity as the outcome variable. For example, one study investigated the amount of daily physical activity across four weeks as a function of self-reported self-control strength in university students intending to exercise regularly (Schondube, Bertrams, Sudeck, & Fuchs, 2017). Low self-reported self-control strength was associated with less physical activity during the day, particularly in individuals low in trait self-control. In another study, the degree to which participants faced situational constraints during their work day that presumably entailed pronounced self-regulatory demands was associated with less physical activity during the same day in a sample of police employees during one work week (Sonnentag & Jelden, 2009). This effect was partially mediated by feelings of low energy. In addition, situational constraints during the work day were associated with lower effort activities later on the same day, such as watching TV.

Some studies operationalized daily self-control demands in close agreement with the ego depletion idea by asking for the extent to which participants had to regulate their mood, control their thoughts, and deal with stress. One study used a sample of college students and employed an experience sampling phase of 49 days that spread out in six chunks over more than one year in total (Simons, Wills, Emery, & Spelman, 2016). Higher daily self-control demands were associated with (a) an increased likelihood of experiencing interpersonal conflict (e.g., getting in verbal arguments) during the same day and (b) neglecting responsibilities such as studying or taking care of daily living responsibilities. In a study on an underage community sample of social drinkers, daily self-control demands were associated with an increased likelihood of violating one’s self-imposed drinking limit (Muraven, Collins, Shiffman, & Paty, 2005). In a sample of employees of various companies, self-control demands were associated with an affect regulation motive, which in turn was associated with an increased intake of sweet snacks at work (Sonnentag, Pundt, & Venz, 2017). Finally, one study tracked experienced desires in various domains (e.g., eating, drinking, smoking, sex, media etc.) in a community sample; and whether participants had tried to resist these desires or not. The more often participants had attempted to resist a desire earlier that day, the more likely they were to enact a subsequent desire, even despite attempts to resist it (Hofmann, Vohs, & Baumeister, 2012; but see O'Connell, Schwartz, & Shiffman, 2008, for conflicting evidence from repeatedly resisting within a single domain).

Interim conclusion. Several field studies from everyday life show effects in line with the general notion of ego depletion. To avoid misconceptions: These studies are correlational, lack experimental manipulations, are open to alternative explanations, are not yet particularly numerous compared to the laboratory depletion literature, and counterevidence may exist to an unknown extent (e.g., O'Connell et al., 2008; Randles, Harlow, & Inzlicht, 2017). Certainly, these studies do not speak to the (non-)existence of laboratory ego depletion effects. The findings are so diverse that—similar to laboratory depletion effects—different psychological processes may contribute to their emergence. In particular, some of these studies might relate to (mental) fatigue (Hockey, 2013) and raise the question whether ego depletion may be a sort of short-term fatigue (Inzlicht & Berkman, 2015). What these studies do, however, is to provide support for the basic theoretical idea of ego depletion in everyday life. They may caution researchers to dismiss this theoretical idea based on doubts about the current state of laboratory ego depletion research.

Summary, Implications, and Outlook

In this article, we examined the merits and faults of both recent critiques suggesting that ego depletion does not exist and six arguments that might be used to defend ego depletion research against this contention. Our benchmark for this analysis was if (a) a proponent of ego depletion would (and should) be convinced by the critiques that ego depletion does not exist, and (b) a skeptic of ego depletion would (and should) be convinced that ego depletion is, in fact, real. Our conclusions are: First, the doubts and criticisms against ego depletion are substantial and challenging. At the same time, none of these critical issues provides conclusive evidence that ego depletion does not exist. Hence, it appears premature to dismiss the phenomenon. Second, none of the arguments provide conclusive evidence for the existence of ego depletion effects beyond reasonable doubt either. This is frustrating to realize. If this is a central conclusion of a review article after two decades of research and several hundred published studies, something must have gone seriously wrong.

Third, from our perspective, the burden of proof has shifted from the skeptics to the proponents who believe ego depletion to be real. We do not see how skeptics could be convinced based on the currently published evidence alone. Instead, new evidence needs to be gathered that meets the highest research standards. It is crucial for psychological science to give a more convincing answer to the question whether self-control wanes over time: The potential real-life implications are too important and the idea has grown too prominent to leave the evidence where it is. Future research needs to improve on various levels including (a) how empirical work is conducted, and (b) its theoretical development. In the following, we will elaborate on some aspects that we believe will increase the conclusiveness of ego depletion research.

Empiricism

Particularly mistrusted claims need particularly convincing evidence. As long as the doors are as widely open to publication bias, *p*-hacking, and underpowered studies as they have been for years, we doubt that researchers skeptical of ego depletion will be convinced by another hundred studies along the lines of the previous several hundred. Adopting principles of open science and increasing statistical power will be crucial for ego depletion research to regain trust. In addition, we suggest adversarial collaborations and finding ways beyond the sequential-task paradigm to investigate the idea of ego depletion.

Open science. In essence, open science proponents call for more transparent science (Munafò et al., 2017). Future work should make all study materials publicly available so other researchers can scrutinize tasks and instructions. The same goes for raw data and analysis scripts. The scientific community should not be left in doubt about any details of the research and whether peculiarities in the data or data-analysis may have unduly influenced the results. One important aspect of open science is pre-registration. Pre-registrations ask researchers to specify – prior to data collection – all measured variables, experimental conditions, hypotheses, stopping rule of data collection, preparation and analysis of the data (e.g., van 't Veer & Giner-Sorolla, 2016). Pre-registration helps to separate exploratory from confirmatory analyses. It is an effective means to prevent many forms of *p*-hacking and to reduce the file drawer, if made public along with the collected data (e.g., on the Open Science Framework).

Statistical power. In addition to implementing open science principles, future ego depletion needs to increase statistical power. Sample size and thereby statistical power has been unacceptably low in extant ego depletion research (Carter et al., 2015; presuming a small-to-medium true average effect). Studies relying on between-participants designs will inevitably require much larger sample sizes than most previous studies. To make this transition less resource-intense, researchers may want to consider sequential stopping procedures. These procedures allow examining the data before the a priori determined minimal sample size has been met (while controlling for type-I error inflation). Sequential stopping procedures are available both within the traditional frequentist statistical approach (Lakens, 2014) and in the Bayesian statistics approach (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). Based on the interim analyses and a priori defined thresholds of evidence researchers may decide whether or not they believe it is necessary or worth it, respectively, to continue data collection. This makes sequential stopping procedures more resource-efficient than mandatorily collecting data up to the a priori determined sample size.

It is often overlooked that statistical power is not only a function of sample size. Instead, power can and should also be enhanced by reducing error variance (e.g., MacKinnon, 2013). This can be done in multiple ways. First, researchers can aim to increase the reliability of their experimental manipulations and outcome measures (e.g., by highly standardizing all study procedures, pretesting questions to confirm participants understand them unambiguously, increasing the number of trials in tasks, etc.). A second way to reduce error is to control for confounding variables that correlate with the outcome variable, but are uncorrelated with the predictor variable(s). For example, researchers studying food intake in an ego depletion study may want to control for self-reported hunger or time since last food intake. These variables are known to influence the amount eaten, but are not of interest for testing the researcher's hypothesis. Finally, a promising approach to reduce error and increase power is by using within-participants designs. For example, researchers could assess a performance baseline of the outcome variable prior to the depletion manipulation and include this baseline as a covariate in the later analyses (Simonsohn, 2015).

Adversarial collaboration. We encourage proponents and critics of ego depletion to team up for (pre-registered) adversarial collaborations in which both sides of the debate agree on particular study setups and make opposing predictions. Data collection may take place in the involved laboratories or even be conducted by third parties. This approach could also be extended to meta-analyses. Existing meta-analyses (Blázquez et al., 2017; Carter et al., 2015; Dang, 2017; Hagger et al., 2010) suffer from several limitations and are disputed by the opposing sides of the debate. In a more encompassing new meta-analysis, members of both sides could jointly determine inclusion criteria, ways to code study quality, strength of manipulations, potential moderators, etc. We are convinced that - independent of their results – such adversarial collaborations have the potential to provide particularly compelling evidence that would improve trust in ego depletion research.

Leaving the lab. Future work should move beyond the typical sequential-task paradigm in the laboratory that has dominated ego depletion research for two decades. Ultimately, (social) psychological research should be able to understand and make predictions about people's lives. The sequential-task paradigm is one way of testing the general idea about ego depletion. It certainly is not the only way and it may not be the best way. To this end, the general idea of ego depletion could be investigated outside the lab using archival data, experience sampling, or ecological momentary interventions that allow implementing experimental manipulations in real life (Heron & Smyth, 2010). Field studies on ego depletion may bear the potential to increase both replicability and real-world impact (Maner, 2016). Initial evidence presented in the section on everyday life evidence of ego depletion suggests that broadening the scope of research and potentially combining it with lab-based research may be a fruitful endeavor.

Theory

Up to this point, we have focused on questions and problems related to how empirical research on ego depletion was conducted. This was a consequence of the central aim of this article to examine the existing literature for conclusive empirical evidence against versus for the existence of ego depletion on the functional level of analysis (i.e., the behavioral effect). We now broaden the scope and consider how these problems relate to and are intertwined with theorizing on ego depletion. We argue that at least part of the replicability problems that may appear to be primarily empirical problems may in fact be also due to insufficient theorizing.

Precision of theory. We contend that theory about ego depletion would profit from more precise and more clearly articulated theoretical assumptions. These should be accompanied by specific suggestions about appropriate measurement that will lead to clearly supporting versus contradicting evidence. To give but one example how a lack of precise theorizing weakens ego depletion research, let us come back to the issue of manipulation checks discussed earlier.

Although there has been debate about how informative manipulation checks really are (Fayant, Sigall, Lemmonier, Retsin, & Alexopoulos, 2017; Sigall & Mills, 1998), many researchers believe them to be very useful or even mandatory tools to provide evidence that the *construct of interest* has in fact been successfully manipulated (Hüffmeier, Mazei, & Schultze, 2016; Stroebe & Strack, 2014). According to the functional definition of ego depletion that we provided at the outset of this article, the exertion of self-control (i.e., control of dominant responses) is the causal antecedent – and therefore the construct of interest – of subsequently impaired self-control performance. Thus, manipulation checks should focus on the amount of control required by a given ego depletion manipulation. This could be done objectively, for example, by specifying the number of dominant responses that participants have to control during an ego depletion manipulation (e.g., number of congruent versus incongruent Stroop trials) or by assessing the exerted effort in the first depleting of a sequential task paradigm (Lee, Chatzisarantis, & Hagger, 2016). Alternatively, researchers could assess self-reports about the (subjective) amount of control or demand that a task requires. While these questions refer to the amount of control required by the first of two consecutive tasks, it is difficult to specify how much control is needed to be exerted for impaired subsequent performance to occur. Put differently, how much control does a manipulation need to entail for a study to provide evidence *against* the hypothesis (if no behavioral impairment is observed)? There are no agreed-upon manipulation checks that have been thoroughly tested and that could be routinely included in ego depletion studies.

One implication is that work on ego depletion has failed to thoroughly validate treatments independent from the behavioral outcomes they produced, perhaps because no theory has clearly articulated validation criteria. The same lack of construct validity may apply to the many dependent variables used in the ego depletion literature. We do not know if and to what extent these dependent variables actually measure the construct we want to measure (i.e., control). In short, because clearly articulated theory is lacking, many constructs in the ego depletion literature—be they independent or dependent variables—have unknown validity (Flake, Pek, & Hehman, 2017; Lurquin & Miyake, 2017).

Coming back to independent variables, in addition to assessing the amount of control they require, typical manipulation checks in ego depletion research also seek to provide evidence for the processes that presumably underlie the behavioral effect. This approach conflates the behavioral, functional level with the process- or cognitive level of analysis (De Houwer, 2011). Mediating processes may be caused by the previous exertion of control (the proposed causal antecedent of the behavioral effect), but they may also be caused by alternative causes not specified in any theoretical hypothesis under investigation (Fiedler, 2017). To be sure, measures of proposed mediating variables can be informative about underlying mechanisms of a behavioral effect. However, manipulation checks should focus on the proposed causal antecedent of the behavioral effect and ideally be independent of the presumed psychological mechanism (De Houwer, 2011).

To provide evidence for an underlying mechanism, researchers would have to assess quite different psychological processes dependent on the theoretical model they want to test. For example, proponents of the strength model of self-control (Baumeister & Vohs, 2016b) would have to provide evidence for diminished self-regulatory *resources*. However, what this resource is, and how its reduction could be measured remains unclear. By contrast, proponents of the opportunity-cost model (Kurzban et al., 2013) would have to measure *cost/benefit computations* associated with working on demanding, often not very exciting tasks. Yet, these computations may be difficult to access introspectively. Proponents of the process model (Inzlicht & Schmeichel, 2012) would have to provide evidence for reduced *motivation* to exert control. While motivation may seem more tangible to assess, even this is challenging due to limits of introspective insight and self-presentation tendencies of participants who may be reluctant to report that they feel like slacking off for the remainder of a study (Brown & Bray, 2017; Silvia & Gendolla, 2001; Wilson & Dunn, 2004). Thus, none of the proposed constructs is easily validly assessed.

In light of these difficulties, researchers regularly ask for subjective downstream consequences of engaging in the first task (e.g., fatigue, frustration, mood, etc.). However, which of these qualities presumably mediate depletion effects is usually not specified in theory; their presumed role in mediating between the actual causal antecedent (i.e., exerted self-control) and the purportedly influenced construct (e.g., resources, motivation) is also unspecified. Major theories agree that effort is important (Baumeister & Vohs, 2016b; Inzlicht & Schmeichel, 2012; Kurzban et al., 2013). But which effort? Is one true mediator objectively exerted effort during the first task? This could be assessed with psychophysiological indicators such as pre-ejection period, systolic blood pressure, or even pupil dilation (Blascovich, Mendes, Vanman, & Dickerson, 2011; Lin, Saunders, Hutcherson, & Inzlicht, in press; Richter, Gendolla, & Wright, 2016). Alternatively, maybe it is subjectively exerted effort that is a key mediator, which does not manifest psychophysiologically. Perhaps depletion is not even about the objective levels of exerted effort, but more about the perceived change in the subjective state associated with having exerted effort (Bless & Burger, 2016). The point here is that for conclusive evidence for a presumed mechanism involving effort, theories need to more clearly specify what they have in mind when they talk about the role effort may play for ego depletion effects (Lurquin & Miyake, 2017).

Bricks versus buildings. In our metaphorical introductory example, we described how the ego depletion house suffers from cracks in walls and floors. Several factors have contributed to this regrettable situation: Neither the quality of the bricks – the individual studies on ego depletion – nor the building as a whole – the theoretical structure – have met ambitious standards. The bulk of the present article was concerned with the quality of individual bricks (studies). No stable building can be built based on bricks that are too small and too instable (Gray, 2017). Therefore, it is imperative to improve the quality of individual studies in ego depletion research, for example, along the lines outlined in the preceding sections. However, even an assembly of highest quality bricks does not make for good science. Good science requires a structure, a plan about how to combine the bricks to a coherent building. Good science gives advice about how to fill gaps between bricks. Only the combination of high quality bricks (studies) and a coherent structure (theory) will advance science in the most efficient way and lead to buildings that are stable and allow for progress in understanding (Gray, 2017). It strikes us that building many new bricks, however loosely connected to previous ones, dominated ego depletion research for too long. The literature is full of evidence about yet another activity that supposedly causes ego depletion effects and about yet another behavior or judgment from various spheres of life that is affected by ego depletion manipulations. This development has blurred the conceptual borders, which has led to an inclusion of tests in the literature that ended up not testing any theory the way it was devised.

Building up this increasingly sizable pile of largely unconnected bricks was not accompanied with similarly impactful contributions about how to integrate the diverse results into a coherent building. Although recent years have seen more efforts at coherent theorizing around ego depletion (e.g., Evans, Boggero, & Segerstrom, 2016; Inzlicht & Schmeichel, 2012; Kurzban et al., 2013), efforts of the field to systematically examine their validity have been too sparse. To be sure, many roads lead to Rome. Building a theory on strong a priori incontestable pillars that allow to deductively arrive at concrete testable hypotheses is regarded by many as the preferred way to conduct high quality science (e.g., Fiedler, 2017). However, this path is not without alternatives. There are also examples of successful inductive theory building where one high quality brick after the other was put together to form a coherent building without knowing from the start what the building would eventually look like (e.g., goal setting theory, Locke & Latham, 2015, or developement of the big five, O. P. John, Naumann, & Soto, 2008). From our perspective, the point is less about deductive versus inductive theorizing, but on prioritizing to systematically build a coherent structure rather than producing large amounts of different bricks with little effort to connect them.

In the course of the replicability crisis, recent years have seen a stark increase in the sensitivity for methodological questions pertaining to power, measurement, data analysis, openness, and transparency. Some researchers have expressed concerns that the pronounced focus on methodological development has led to a disregard of the importance of theory for scientific progress (Fiedler, 2017; Gray, 2017; Schwarz & Clore, 2016; Stroebe & Strack, 2014). Even the largest, most rigorous, open and transparent study will fail to appreciably advance science if it is grounded in poor theory. We agree with this contention. Neither methodological nor theoretical developments are sufficient conditions for ego depletion research, or psychological science more generally for that matter, to thrive. Both are necessary conditions, should be pursued, and should not be played off against each other.

Implications for other fields

Although the present article has focused on ego depletion research, several of the problems that plague this field are similar in other challenged fields (i.e., publication bias, *p*-hacking, poor bricks and not enough focus on buildings). Consequently, we believe that the suggestions outlined in the previous sections are broadly applicable to those fields and would similarly strengthen the conclusiveness of the respective literatures. Instead of reiterating these suggestions here, we would like to focus on one additional issue of crucial importance not only for ego depletion research, but research in psychology in general.

In this article, we have discussed *p*-hacking and publication bias as two critical problems with severe implications for the replicability of the literature and the trust that the scientific community and the public have in this literature. More generally, there is a broad consensus that these two issues constitute serious threats to psychological science (Franco et al., 2016; Munafò et al., 2017). What has been discussed less is the relative impact of publication bias and *p*-hacking. How severe are the problems that are caused by each of the two, and how do they work together to bias the conclusions that researchers draw from a given literature? There are different ways to address this issue. One way to discuss this is to look at the relative potency of publication bias and *p*-hacking for distorting meta-analytic effect size estimates. Meta-analyses are regarded the most suitable method to develop an impression of mean quantitative relationships and their moderating conditions in psychological science, but meta-analyses are plagued by the systematic biases in the published literature that are caused by publication bias and *p*-hacking (Carter et al., 2017).

A brief look at Figure 1A provides a first idea about the relative impact of publication bias and *p*-hacking on the estimate of the mean ego depletion effect. Recall that Figure 1A is a hypothetical sample of 1000 studies with a true effect of zero. The black dots outside of the funnel to the right are studies that produced a significant depletion effect in the expected direction (genuine false positives). The white dots represent studies that are 'in danger of being *p*-hacked'—they are close enough to significance to make them significant by applying questionable research practices (see footnote 4 for details). It is readily conceivable that if researchers were aware of all studies conducted, then *p*-hacking even the majority of the white-dot studies to significance would hardly change the big picture. There are just too many black-dot studies for the relatively few white-dot studies to make an appreciable impact. This is easy to quantify by fitting a simple, fixed effects meta-analytic model to the effect sizes depicted in the figure. The summary effect for all effect sizes in Figure 1A is exactly $d = 0.006$. This slight deviation from zero is owed to the fact that studies were randomly sampled. In Figure 1B, about 50% of the white dots ($k = 75$ studies of all studies identified as 'in danger of being *p*-hacked') were actually *p*-hacked to significance. Nevertheless, even after *p*-hacking, the true effect size in Figure 1B is only $d = 0.022$. Thus, in this example, *p*-hacking led to a shift in effect size of only $\Delta d = 0.016$.

Now, in reality, meta-analysts never have access to all studies that have been conducted. Going back to Figure 1A, suppose that most of the black-dot studies outside of the funnel to the right are published, but many of the other (non-significant) black-dot studies remain unpublished in the file drawer (publication bias). As a result, the meta-analytic (over-) estimate would be strongly influenced by the black-dot studies outside the funnel. In this situation, *p*-hacking (and publishing) the majority of the white-dot studies could exert a stronger impact on the meta-analytic effect size estimate. This is because many more studies would now provide 'evidence' for the hypothesized effect without a similarly strong counter weight of non-significant black-dot studies as in the previous scenario. Thus, the impact of the significant studies (genuine false positives plus *p*-hacked studies) for the meta-analytic effect size estimate would increase. Taken together, a preliminary hypothesis based on the visual inspection of Figure 1A alone is that (a) publication bias always leads to severe overestimations of the meta-analytic effect size, and (b) *p*-hacking leads to only modest bias in the effect size estimate in the (unrealistic) case that all studies are known; but *p*-hacking works together with publication bias when predominantly published, significant studies are available to the meta-analyst.

To address this issue more formally, we drew on recent simulation work that more generally investigates the relative potency of *p*-hacking and publication bias to distort mean effect size estimates in psychological research literatures (Friese & Frankenbach, 2018). For present purposes, we adapted the general approach taken in this work to the specifics of the ego depletion literature, as defined by the parameters that went into Figure 1. Let us briefly reiterate these parameters: We simulated 1,000 studies with a true, fixed effect size of $d = 0$ and no heterogeneity, using empirical sample sizes from the ego depletion literature (see Figure 1A). Then, we simulated *p*-hacking by (a) specifying a range of non-significant *p*-values that can realistically be hacked (footnote 4), and (b) actually hacking these effects with a probability of 50%.

In the following simulation, we additionally realized conditions in which 0% and 100%, respectively, of the studies identified as being 'in danger of being *p*-hacked' were in fact hacked. To operationalize several degrees of publication bias, we specified that a random sample of 100% (highest degree of publication bias), 95%, 70%, or 0% (no publication bias) of the studies that are not significant in the predicted direction are removed from the set (i.e. end up in the file-drawer). By crossing these specifications of *p*-hacking severity and publication bias, we obtained twelve unique combinations. We then applied each *p*-hacking/ publication bias combination to simulated sets of effect sizes that were generated using the exact same procedures that generated the set depicted in Figure 1. Finally, we performed a meta-analysis to examine how the respective *p*-hacking/publication bias combination biased the summary effect estimate.[7]

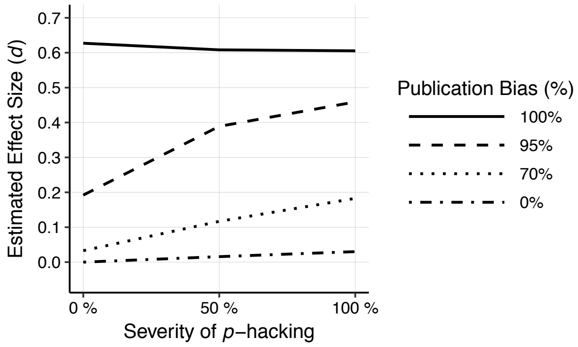

Figure 2. Graphical representation of the relative potency of publication bias and *p*-hacking for biasing meta-analytic fixed-effect summary effects. Estimated (biased) summary effects in the metric of Cohen's *d* are mapped to the *y*-axis. The three categories on the *x*-axis represent three levels of *p*-hacking severity, where either no, half, or all studies identified as 'in danger of being *p*-hacked' are hacked to significance (see Figure 1 and footnote 4). The different lines represent different levels of publication bias, defined by the percentage of non-significant studies that are file-drawered. The true effect size is set to $d = 0$.

The results are depicted in Figure 2. They reveal a few interesting insights: First, in the (unrealistic) scenario of no publication bias, even the most severe form of *p*-hacking hardly biases the meta-analytic effect size estimate (dotted/dashed line at the bottom of Figure 2). In other words: In a world without publication bias, *p*-hacking is not a severe problem (at least not as long as distortion of meta-analytic effect size estimates is taken as the criterion). Second, when publication bias is greatest—no non-significant studies are published at all (continuous black line in Figure 2)—publication bias exerts a strong biasing impact, but adding *p*-hacking does no additional harm. Again, the funnel plot in Figure 1B helps to understand this result: Imagine all black and white dots outside the funnel to the right are averaged (but no further studies). In this case, the white dots (the *p*-hacked studies) provide similar information as the black dots (the genuine false positives). *P*-hacking is therefore unlikely to cause an upward bias to the average of the black dots (i.e., summary effect). Third, publication bias and *p*-hacking interact. If 95% of the non-significant effects are file-drawered, going from 0% to 100% *p*-hacking introduces an upward bias of $\Delta d = 0.27$ (dashed line). However, if only 70% of the non-significant effects are file-drawered, the effect bias due to the addition of *p*-hacking is reduced to $\Delta d = 0.15$ (dotted line).

Fourth, even modest reductions of the file-drawer will greatly reduce the upward bias in meta-analytic effect size estimates. If 50% of studies 'in danger of being *p*-hacked' are in fact hacked, and all non-significant studies are file-drawered, the resulting summary effect is a sizable $d = 0.61$ (all bias, because the true effect is zero in this scenario). Remarkably, if only 5% of the non-significant effects are published (i.e., 95% publication bias), the resulting summary effect goes down to $d = 0.39$, a bias reduction of $\Delta d = 0.22$. Going down even further from 95% publication bias to 70%, the resulting summary effect is a meager $d = 0.12$, although still more than two thirds (!) of non-significant studies end up in the file-drawer.

What are the substantial implications of these simulations not only for ego depletion research, but for other fields as well? In the past several years, psychology has seen a surge of new methods and suggested procedures to deal with the challenges that the field is facing. A lot of change for the better has been instigated in a comparatively short period of time (Nelson et al., 2018, see also the above section on empiricism). A particular focus lay on pre-registrations as an easy and effective means to deal with *p*-hacking (Nosek, Ebersole, DeHaven, & Mellor, 2017; van 't Veer & Giner-Sorolla, 2016).

It is our impression that the second major threat that we discussed in this manuscript—publication bias—has not yet been addressed with the same decisiveness. The present simulations suggest that under many circumstances publication bias may have even more severe consequences than *p*-hacking. One clear implication therefore is that the field should go to great lengths to reduce publication bias by increasing the chances that non-significant studies will be published. For example, more journals could pick up the model of registered reports where authors go through a peer-review process and receive an in-principle acceptance before the data have been collected (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015; Jonas & Cesario, 2016). In the traditional manuscript format without registered reports, editors may want to more consistently require authors to transparently report all studies that have been conducted in a series of studies addressing the same research question (Lakens & Etz, 2017). These are very easy, low-cost, and easy-to-implement strategies with great leverage as even a small amount of publishing non-significant studies that will greatly reduce bias. As the treatment of publication bias receives even more attention than it currently does, these measures will be further developed and complemented with new measures.

Admittedly, these suggestions are at odds with recent suggestions that "The file-drawer problem is unfixable, and that's OK" (Simonsohn, 2016). From this latter perspective, it is untenable to publish papers on true null-effects anyway, because no researcher could afford building up a file-drawer large enough to produce a multi-study paper on a (truly non-existing) effect. Plus, study registries intended to remedy the publication bias problem do not work because researchers do not reliably share their results even when studies have been pre-registered (e.g., Anderson et al., 2015; Franco et al., 2014). We agree that it will be impossible to prevent a file-drawer altogether. However, the reassuring

[7] All summary effects were computed using fixed-effect meta-analysis models. In order to ensure stability of the estimates, all simulations were run 1,000 times and the results were averaged. The R code can be found at https://osf.io/s79fe/.

implication of the present simulations is that even modest reductions in the size of the file drawer will lead to great reductions in meta-analytic effect size bias.

As with the previous simulations, we hasten to say that these results obviously greatly depend on the specific parameters that we used and the assumptions that we made. For example, we ran these simulations based on empirical sample sizes in the ego depletion literature. In other literatures, the effect may be different. In addition to the specific sample size distribution employed here, other researchers may well disagree how easy or difficult it is to *p*-hack a study, thereby changing the number of studies 'in danger of being *p*-hacked' (see footnote 4). Some researchers expressed their concern that nearly all studies, independent of their original result, can and will be 'successfully' *p*-hacked (Nelson et al., 2018). By contrast, we assumed that there are moral and practical limits to *p*-hacking, especially for those studies that cannot be *p*-hacked in any direction (see footnote 4). For example, in an ego depletion study, *p*-hacking a result when the descriptive difference runs counter the expected depletion effect seems unfeasible. Another potential limitation is that we assumed that there is no effect size heterogeneity in the fixed null effect. In truth, observed effect sizes may vary even when the true ego depletion effect is zero, for example, due to experimenter or demand effects, or any inadvertent and potentially unknown effect on subsequent performance caused by the manipulation (e.g., interest in the study, self-licensing). All this being said, we believe that the general conclusion that even a relatively small amount of published non-significant findings may greatly reduce to degree of bias in meta-analytic effect size estimates could be quite robust.

To us, this latter notion is somewhat reassuring, along with some other encouraging developments (Nelson et al., 2018). There seem to be easy, low-cost levers to deal with two of the most pressing problems our field is facing: *P*-hacking and publication bias. Publishing even small random samples of non-significant findings may greatly reduce bias in meta-analytic effect size estimates. Pre-registering studies is easily done and greatly reduces *p*-hacking. Fortunately, pre-registration of studies becomes more common also in ego depletion research (Alquist et al., 2018; Dang, Liu, Liu, & Mao, 2017; Lurquin et al., 2016). In addition, researchers have started to address some of the other issues we discussed in this article: A thorough validation of basic assumptions and some commonly-used measures has started (Arber et al., 2017; Goldberg et al., 2017). The importance of theory is increasingly recognized, and theory-guided work becomes more visible (Dang, Bjorklund, et al., 2017; Lurquin & Miyake, 2017; Molden, Hui, & Scholer, 2017). It is disturbing that the question whether or not ego depletion exists has to be called an open question. But at least there is some hope that future work will not take another 20 years to generate more conclusive evidence to answer this question, one way or the other.

Conclusion

Whether or not ego depletion is real is subject to great debate. Our analysis suggests that the critical evidence is unlikely to convince proponents that ego depletion does not exist. Likewise, the supporting evidence is unlikely to convince skeptics that ego depletion does exist. Better empiricism and better theory are needed to move the field forward and find more conclusive answers to the question if, when, and why ego depletion does (not) exist.

References

Agrawal, N., & Wan, E. W. (2009). Regulating risk or risking regulation? Construal levels and depletion effects in the processing of health messages. *Journal of Consumer Research, 36*, 448-462.

Alquist, J. L., Baumeister, R. F., McGregor, I., Core, T. J., Benjamin, I., & Tice, D. M. (2018). Personal conflict impairs performance on an unrelated self-control task: Lingering costs of uncertainty and conflict. *Journal of Experimental Social Psychology, 74*, 157-160.

Anderson, M. L., Chiswell, K., Peterson, E. D., Tasneem, A., Topping, J., & Califf, R. M. (2015). Compliance with results reporting at ClinicalTrials.gov. *New England Journal of Medicine, 372*, 1031-1039.

Arber, M. M., Ireland, M. J., Feger, R., Marrington, J., Tehan, J., & Tehan, G. (2017). Ego depletion in real-time: An examination of the sequential-task paradigm. *Frontiers in Psychology, 8*.

Augenblick, N., & Nicholson, S. (2016). Ballot position, choice fatigue, and voter behaviour. *Review of Economic Studies, 83*, 460-480.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Baumeister, R. F. (2014). Self-regulation, ego depletion, and inhibition. *Neuropsychologia, 65*, 313-319.

Baumeister, R. F., Bratslavsky, M., 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., Gailliot, M. T., De Wall, C. N., & Oaten, M. (2006). Self-regulation and personality: How interventions increase regulatory success, and how depletion moderates the effects of traits on behavior. *Journal of Personality, 74*, 1773-1801.

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

Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1994). *Losing control: How and why people fail at self-regulation*. San Diego, CA: Academic Press.

Baumeister, R. F., & Tierney, J. (2011). *Willpower. Rediscovering the greatest human strength*. New York: Penguin.

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

Baumeister, R. F., & Vohs, K. D. (2016b). Strength model of self-regulation as limited resource: Assessment, controversies, update. In M. O. James & P. Z. Mark (Eds.), *Advances in Experimental Social Psychology* (Vol. 54, pp. 67-127). San Diego, CA: Academic Press.

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

Beames, J. R., Schofield, T. P., & Denson, T. F. (2018). A meta-analysis of improving self-control with practice. In D. T. D. de Ridder, M. A. Adriaanse & K. Fujita (Eds.), *Handbook of self-control in health and well-being* (pp. 405-417). Abingdon, UK: Routledge.

Blascovich, J., Mendes, W. B., Vanman, E., & Dickerson, S. (2011). *Social psychophysiology for social and personality psychology*. London: Sage.

Blázquez, D., Botella, J., & Suero, M. (2017). The debate on the ego-depletion effect: Evidence from meta-analysis with the p-uniform method. *Frontiers in Psychology, 8*, 197.

Bless, H., & Burger, A. M. (2016). A closer look at social psychologists' silver bullet: Inevitable and evitable side effects of the experimental approach. *Perspectives on Psychological Science, 11*, 296-308.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Brown, D. M. Y., & Bray, S. R. (2017). Graded increases in cognitive control exertion reveal a threshold effect on subsequent physical performance. *Sport Exercise and Performance Psychology, 6*, 355-369.

Bush, G., Shin, L. M., Holmes, J., Rosen, B. R., & Vogt, B. A. (2003). The Multi-Source Interference Task: validation study with fMRI in individual subjects. *Molecular Psychiatry, 8*, 60-70.

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

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

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2017). *Correcting for bias in psychology: A comparison of meta-analytic methods*. Retrieved from osf.io/rf3ys

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex, 66*, A1-A2.

Chow, J. T., Hui, C. M., & Lau, S. (2015). A depleted mind feels inefficacious: Ego-depletion reduces self-efficacy to exert further self-control. *European Journal of Social Psychology, 45*, 754-768.

Christian, M. S., & Ellis, A. P. J. (2011). Examining the effects of sleep deprivation on workplace deviance: A self-regulatory perspective. *Academy of Management Journal, 54*, 913-934.

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*, 118-128.

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

Cunningham, M. R., & Baumeister, R. F. (2016). How to make nothing out of something: Analyses of the impact of study sampling and statistical interpretation in misleading meta-analytic conclusions. *Frontiers in Psychology, 7*.

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

Dang, J. H. (2016). Commentary: A multilab preregistered replication of the ego depletion effect. *Frontiers in Psychology, 7*.

Dang, J. H. (2017). An updated meta-analysis of the ego depletion effect. *Psychological Research*.

Dang, J. H., Bjorklund, F., & Backstrom, M. (2017). Self-control depletion impairs goal maintenance: A meta-analysis. *Scandinavian Journal of Psychology, 58*, 284-293.

Dang, J. H., Liu, Y., Liu, X., & Mao, L. (2017). The ego could be depleted, providing initial exertion is depleting. *Social Psychology, 48*, 242-245.

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science, 6*, 202-209.

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.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality, 45*, 259-268.

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

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*, 572-583.

Etherton, J. L., Osborne, R., Stephenson, K., Grace, M., Jones, C., & De Nadai, A. (in press). Bayesian analysis of multimethod ego-depletion studies favours the null hypothesis. *British Journal of Social Psychology*.

Evans, D. R., Boggero, I. A., & Segerstrom, S. C. (2016). The nature of self -regulatory fatigue and "ego depletion": Lessons from physical fatigue. *Personality and Social Psychology Review, 20*, 291-310.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *Plos One, 5*.

Fayant, M. P., Sigall, H., Lemmonier, A., Retsin, E., & Alexopoulos, T. (2017). On the limitations of manipulation checks: An obstacle toward cumulative science. *International Review of Social Psychology, 30*, 125-130.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12*, 46-61.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*, 45-52.

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

Fischer, P., Greitemeyer, T., & Frey, D. (2007). Ego depletion and positive illusions: Does the construction of positivity require regulatory resources? *Personality and Social Psychology Bulletin, 33*, 1306-1321.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*, 370-378.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502-1505.

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science, 7*, 8-12.

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

Friese, M., & Frankenbach, J. (2018). *P-hacking versus publication bias: Which is the bigger threat?*, Unpublished manuscript.

Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science, 12*, 1077-1099.

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

Goldberg, S. B., Flook, L., Hirshberg, M. J., Findley, D., Kesebir, P., Schaefer, S. M., et al. (2017). Getting a grip on the handgrip task: Handgrip duration correlates with neuroticism but not conscientiousness. *Frontiers in Psychology, 8*.

Govorun, O., & Payne, B. K. (2006). Ego depletion and prejudice: Separating automatic and controlled components. *Social Cognition, 24*, 111-136.

Graham, J. D., Ginis, K. A. M., & Bray, S. R. (2017). Exertion of self-control Increases fatigue, reduces task self-efficacy, and impairs performance of resistance exercise. *Sport Exercise and Performance Psychology, 6*, 70-88.

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science, 12*, 731-741.

Gröpel, P., Baumeister, R. F., & Beckmann, J. (2014). Action versus state orientation and self-control performance after depletion. *Personality and Social Psychology Bulletin, 40*, 476-487.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., et al. (2016). A multi-lab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546-573.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136*, 495-525.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York, NY: Guilford Press.

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*, 573-585.

Heatherton, T. F., & Wagner, D. D. (2011). Cognitive neuroscience of self-regulation failure. *Trends in Cognitive Sciences, 15*, 132-139.

Heckman, B. W., MacQueen, D. A., Marquinez, N. S., MacKillop, J., Bickel, W. K., & Brandon, T. H. (2017). Self-control depletion and nicotine deprivation as precipitants of smoking cessation failure: A human laboratory model. *Journal of Consulting and Clinical Psychology, 85*, 381-396.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61-85.

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology, 15*, 1-39.

Hirt, E. R., Clarkson, J. J., & Jia, L. (2016). *Self-regulation and ego control*. San Diego, CA: Elsevier.

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

Hofmann, W., Vohs, K. D., & Baumeister, R. F. (2012). What people desire, feel conflicted about, and try to resist in everyday life. *Psychological Science, 23*, 582-588.

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology, 66*, 81-92.

Imhoff, R., Schmidt, A. F., & Gerstenberg, F. (2014). Exploring the interplay of trait self-control and ego depletion: Empirical evidence for ironic effects. *European Journal of Personality, 28*, 413-424.

Inzlicht, M., & Berkman, E. (2015). Six questions for the resource model of control (and some answers). *Social and Personality Psychology Compass, 9*, 511-524.

Inzlicht, M., Gervais, W. M., & Berkman, E. T. (2015). Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015. *Available at SSRN: http://ssrn.com/abstract=2659409*.

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

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

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

Job, V., Dweck, C. S., & Walton, G. M. (2010). Ego depletion-Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological Science, 21*, 1686-1693.

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

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology, 1*, 1-7.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology-General, 143*, 131-141.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences, 36*, 661-679.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*, 701-710.

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science, 8*, 875-881.

Lakens, D., Hilgard, J., & Staaks, J. J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*, 24.

Lee, N., Chatzisarantis, N., & Hagger, M. S. (2016). Adequacy of the sequential-task paradigm in evoking ego-depletion and how to improve detection of ego-depleting phenomena. *Frontiers in Psychology, 7*.

Levav, J., Heitmann, M., Herrmann, A., & Iyengar, S. S. (2010). Order in product customization decisions: Evidence from field experiments. *Journal of Political Economy, 118*, 274-299.

Lewis, M. (2012). Obama's way. *Vanity Fair*. Retrieved from http://www.vanityfair.com/news/2012/10/michael-lewis-profile-barack-obama

Lin, H., Saunders, B., Hutcherson, C. A., & Inzlicht, M. (in press). Midfrontal theta and pupil dilation parametrically track subjective conflict (but also surprise) during intertemporal choice. *NeuroImage*.

Linder, J. A., Doctor, J. N., Friedberg, M. W., Nieva, H. R., Birks, C., Meeker, D., et al. (2014). Time of day and the decision to prescribe antibiotics. *Jama Internal Medicine, 174*, 2029-2031.

Locke, E. A., & Latham, G. P. (2015). Breaking the rules: A historical overview of goal-setting theory. In A. J. Elliot (Ed.), *Advances in Motivation Science* (Vol. 2, pp. 99-126): Elsevier.

Loschelder, D. D., & Friese, M. (2016). Moderators of the ego depletion effect. In E. R. Hirt, J. J. Clarkson & L. Jia (Eds.), *Self-regulation and ego control* (pp. 21-42). San Diego, CA: Elsevier.

Luethi, M., Friese, M., Binder, J., Boesiger, P., Luechinger, R., & Rasch, B. (2016). Motivational incentives lead to a strong increase in lateral prefrontal activity after self-control exertion. *Social Cognitive and Affective Neuroscience, 11*, 1618-1626.

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., von Bastian, C. C., Carruth, N. P., et al. (2016). No evidence of the ego-depletion effect across task characteristics and individual differences: A pre-registered study. *Plos One, 11*.

Lurquin, j. H., & Miyake, A. (2017). Challenges to ego-depletion research go beyond the replication crisis: A need for tackling the conceptual crisis. *Frontiers in Psychology, 8*.

MacKinnon, S. (2013, October 13). Increasing statistical power in psychological research without increasing sample size. Retrieved from http://osc.centerforopenscience.org/2013/11/03/Increasing-statistical-power/

Maner, J. K. (2016). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology, 66*, 100-106.

McGonigal, K. (2012). *The willpower instinct: How self-control works, why it matters, and what you can do to get more of it*. New York, NY: Penguin.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*, 730-749.

Molden, D. C., Hui, C. M., & Scholer, A. A. (2017). What limits self-control? A motivated effort-allocation account. In D. T. D. De Ridder, M. A. Adriaanse & K. Fujita (Eds.), *The Routledge handbook of self-control in health and well-being* (pp. 129-142). London: Routledge.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., et al. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *Bmc Medical Research Methodology, 9*, 2.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *1*, 0021.

Muraven, M., & 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*, 113-120.

Muraven, M., Collins, R. L., Shiffman, S., & Paty, J. A. (2005). Daily fluctuations in self-control demands and alcohol intake. *Psychology of Addictive Behaviors, 19*, 140-147.

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

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

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18*, 107-118.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2017). *The preregistration revolution.* Retrieved November 23, 2017, from osf.io/2dxu5

O'Connell, K. A., Schwartz, J. E., & Shiffman, S. (2008). Do resisted temptations during smoking cessation deplete or augment self-control resources? *Psychology of Addictive Behaviors, 22*, 486-495.

Osgood, J. M. (2017). Effect of ego-depletion typing task on Stroop does not extend to diverse online sample. *Journal of Articles in Support of the Null Hypothesis, 13*, 83-89.

Pocheptsova, A., Amir, O., Dhar, R., & Baumeister, R. F. (2009). Deciding without resources: Resource depletion and choice in context. *Journal of Marketing Research, 46*, 344-355.

Randles, D., Harlow, I., & Inzlicht, M. (2017). A pre-registered naturalistic observation of within domain mental fatigue and domain-general depletion of self-control. *Plos One, 12*, e0182980.

Richard, F. D., Bond, C. F., & Stokes-Zoota. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363.

Richter, M., Gendolla, G. H. E., & Wright, R. A. (2016). Three decades of research on Motivational Intensity Theory: What we have learned about effort and what we still don't know. In A. J. Elliot (Ed.), *Advances in Motivation Science* (Vol. 3, pp. 149-186): Elsevier.

Robinson, M. D., Schmeichel, B. J., & Inzlicht, M. (2010). A cognitive control perspective of self-control strength and its depletion. *Social and Personality Psychology Compass, 4*, 189-200.

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

Savani, K., & Job, V. (2017). Reverse ego-eepletion: Acts of self-control can improve subsequent performance in Indian cultural contexts. *Journal of Personality and Social Psychology, 113*, 589-607.

Scargle, J. D. (2000). Publication bias: The “file-drawer” problem in scientific inference. *Journal of Scientific Exploration, 14*, 91-106.

Schaller, M. (2016). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology, 66*, 107-115.

Schmeichel, B. J., & Vohs, K. D. (2009). Self-affirmation and self-control: Affirming core values counteracts ego depletion. *Journal of Personality and Social Psychology, 96*, 770-782.

Schönbrodt, F. D. (2015). *p-hacker: Train your p-hacking skills!* Retrieved November 15, 2016, from http://shinyapps.org/apps/p-hacker/

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*, 322-339.

Schondube, A., Bertrams, A., Sudeck, G., & Fuchs, R. (2017). Self-control strength and physical exercise: An ecological momentary assessment study. *Psychology of Sport and Exercise, 29*, 19-26.

Schwarz, N., & Clore, G. L. (2016). Evaluating psychological research requires more than attention to the N. *Psychological Science, 27*, 1407-1409.

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

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

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

Sievertsen, H. H., Gino, F., & Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences of the United States of America, 113*, 2621-2624.

Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review, 2*, 218-226.

Silvia, P. J., & Gendolla, G. H. E. (2001). On introspection and self-perception: Does self-focused attention enable accurate self-knowledge? *Review of General Psychology, 5*, 241-269.

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

Simons, J. S., Wills, T. A., Emery, N. N., & Spelman, P. J. (2016). Keep calm and carry on: Maintaining self-control when intoxicated, upset, or depleted. *Cognition & Emotion, 30*, 1415-1429.

Simonsohn, U. (2014, October 13). [17] No-way interactions. Retrieved from http://datacolada.org/17

Simonsohn, U. (2015, October 13). [39] Power naps: When do within-subject comparisons help vs hurt (yes, hurt) power? Retrieved from http://datacolada.org/39

Simonsohn, U. (2016, November 20). [55] The file-drawer problem is unfixable, and that's OK. Retrieved from http://datacolada.org/55

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-681.

Sonnentag, S., & Jelden, S. (2009). Job stressors and the pursuit of sport activities: A day-level perspective. *Journal of Occupational Health Psychology, 14*, 165-181.

Sonnentag, S., Pundt, A., & Venz, L. (2017). Distal and proximal predictors of snacking at work: A daily-survey study. *Journal of Applied Psychology, 102*, 151-162.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*, 845-851.

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

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*, 60-78.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*, 1046-1055.

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 342*, d4002.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59-71.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

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

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*, 2113-2126.

Tice, D. M., Baumeister, R. F., Shmueli, D., & Muraven, M. (2007). Restoring the self: Positive affect helps improve self-regulation following ego depletion. *Journal of Experimental Social Psychology, 43*, 379-384.

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

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

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*, 293-309.

Van Reet, J. (2015). Conflict inhibitory control facilitates pretense quality in young preschoolers. *Journal of Cognition and Development, 16*, 333-350.

Vohs, K. D., Baumeister, R. F., & Schmeichel, B. J. (2013). Motivation, personal beliefs, and limited resources all contribute to self-control. *Journal of Experimental Social Psychology, 49*, 183-188.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2008). Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *Journal of Personality and Social Psychology, 94*, 883-898.

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

Wan, E. W., & Sternthal, B. (2008). Regulating the effects of depletion through monitoring. *Personality and Social Psychology Bulletin, 34*, 32-46.

Webb, T. L., & Sheeran, P. (2003). Can implementation intentions help to overcome ego-depletion? *Journal of Experimental Social Psychology, 39*, 279-286.

Wenzel, M., Kubiak, T., & Conner, T. S. (2014). Positive affect and self-control: Attention to self-control demands mediates the influence of positive affect on consecutive self-control. *Cognition & Emotion, 28*, 747-755.

Wilson, T. D., & Dunn, E. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology, 55*, 493-518.

Xu, X. M., Demos, K. E., Leahey, T. M., Hart, C. N., Trautvetter, J., Coward, P., et al. (2014). Failure to replicate depletion of self-control. *Plos One, 9*, e109950.

Yam, K. C., Chen, X. P., & Reynolds, S. J. (2014). Ego depletion and its paradoxical effects on ethical decision making. *Organizational Behavior and Human Decision Processes, 124*, 204-214.

Yost, T. A. (2016). *Meta-analytic and empirical estimates of the resource depletion effect size*. University of Minnesota Digital Conservancy. Retrieved from http://hdl.handle.net/11299/181688