A Multi-Site Preregistered, Paradigmatic Test of the Ego Depletion Effect

SUPPLEMENTARY ONLINE MATERIALS

Exploratory Analyses

Depletion effect

Frequentist analyses. Analyses based on the full dataset were not preregistered, but the rate of exclusions far exceeded expectations. We therefore decide to conduct exploratory analyses using the full dataset.

Meta-analyses of the full dataset revealed a small significant effect in line with predictions (RE: d = 0.08, 95% CI [0.01, 0.15]; FE: d = 0.07, 95% CI [0.01, 0.14]; I^2 = 11.69%; Figure S1). This effect was observed for both random- and fixed-effects models. Experimental protocol did not appear to moderate the depletion effect, RE: intercept d = 0.08 [0.00, 0.15], moderator b = -0.07 95% CI [-0.22, 0.07], I^2 = 13.90%.

We also tested whether there was evidence of an overall depletion effect using multilevel regression approaches that nested the individual-level data within laboratories in random-intercept mixed models. In the reduced sample (excluding 1068 participants, following preregistered rules), task performance did not differ by depletion condition, b = 0.09 CI [-0.01, 0.19]. In the full sample (when participants marked for exclusion were included), the effect of depletion condition was statistically significant but small (Table S1).

Figure S1. *Forest Plot of Performance Outcome by Laboratory: Full Sample.* The box plots and numerical values illustrate the same effect size estimates. For the plots, the size of the box represents its weighted contribution to the overall effect and its whiskers display 95% CIs. The dotted line represents a zero effect size. Numerical values show standardized mean differences between depletion and non-depletion conditions expressed in Cohen's *d* (with 95% CIs). The diamond is the overall meta-analytic effect derived from a random-effects model.

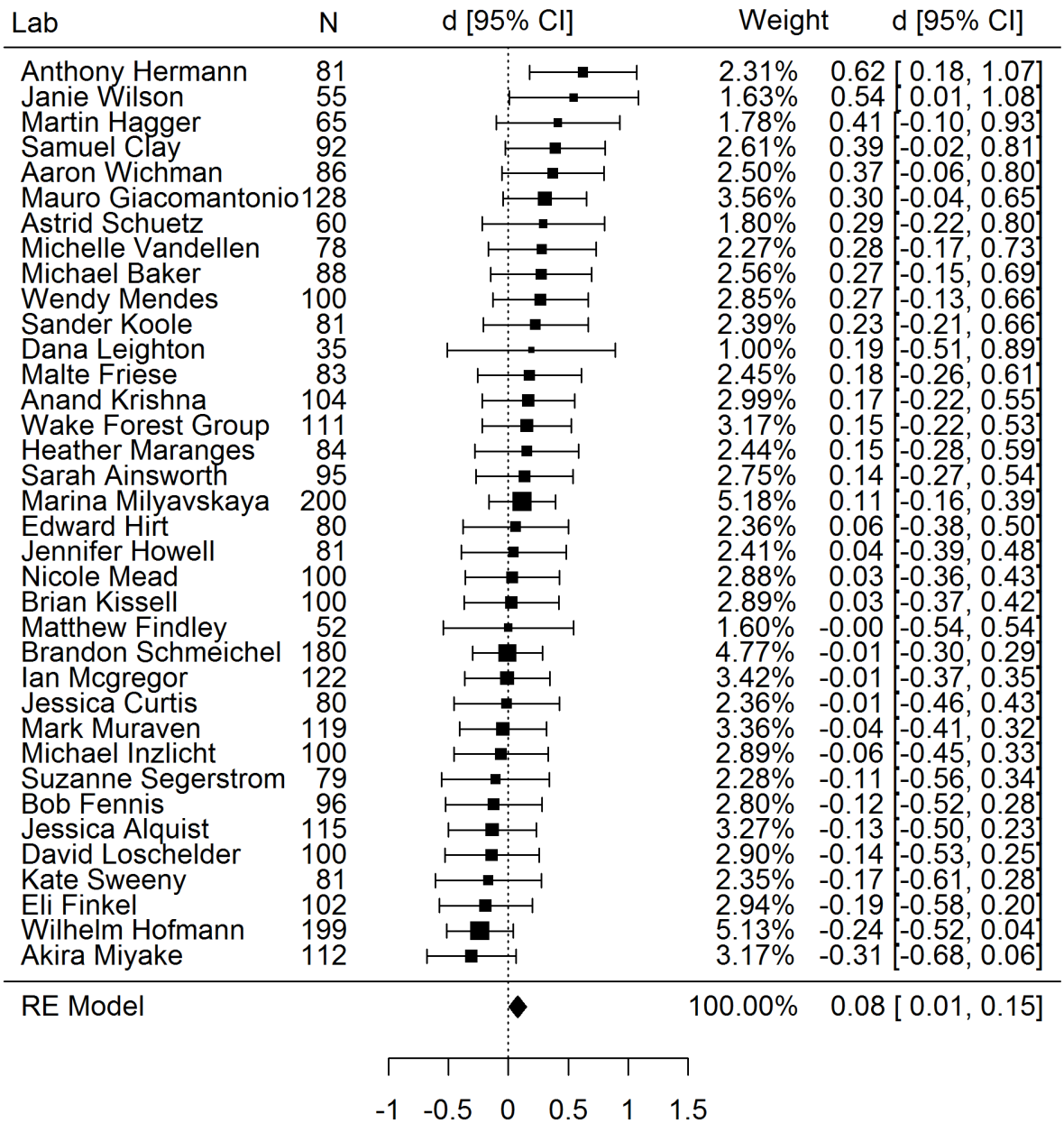


Table S1. *Depletion Effect: Exploratory Frequentist Meta-Analyses and Multi-Level Models*

		Random-effects meta-analysis			Fixed-effects meta-analysis		Multi-level regression	
DV	***N***	***d***	**CI**	**I^2 %**	***d***	**CI**	***b***	**CI**
Overall depletion effect	3524	0.08 *	[0.01, 0.15]	11.69	0.07 *	[0.01, 0.14]	0.11 *	[0.02, 0.20]
Overall figure tracing performance	1847	0.12 *	[0.01, 0.23]	27.23	0.10 *	[0.01, 0.20]	0.18 *	[0.03, 0.32]
Figure tracing duration	1847	0.14 *	[0.01, 0.27]	46.83	0.12 *	[0.03, 0.21]	0.11 *	[0.02, 0.20]
Figure tracing attempts	1848	0.06	[-0.04, 0.15]	0	0.06	[-0.04, 0.15]	0.07	[-0.02, 0.15]
Cognitive Estimation Test	1677	0.04	[-0.06, 0.13]	0	0.04	[-0.06, 0.13]	0.04	[-0.06, 0.13]

Note: Results pertain to the entire sample. Sample sizes vary due to missing data. For overall depletion effect analyses, $k = 36$; figure tracing analyses, $k = 20$; Cognitive Estimation Test analyses, $k = 16$. Condition coded such that 0 = non-depletion, 1 = depletion condition. Higher numbers indicate evidence of a depletion effect (i.e., self-control was worse in the depletion condition). DV stands for dependent variable. FE indicates fixed-effects models; RE indicates random-effects models. *CI* indicates 95% confidence intervals. Multi-level models nested participants' data within labs and used a random intercept for labs. * $p < .05$

Bayesian analyses. We next turn to the model-averaged meta-analytic Bayes factor (which corresponds closely to the fixed- and random-effects Bayes factors; Figure S2). The results indicated that the data are 1.33 times more likely under the point-null hypothesis (which states that the effect is absent) than under the one-sided informed alternative hypothesis (which states that the effect is present), suggesting that two models predict the data almost equally well. Although the full sample data provided no basis for shifting beliefs towards or away from either hypothesis, the posterior distribution addressed the magnitude of the effect if it is present.

To take into account the findings from all laboratories simultaneously, we considered the results of the model-averaged meta-analysis. Figure S3 displays the model-averaged meta-analytic posterior for effect size as a solid line; the dotted line indicates the informed prior distribution. As shown, the data have shifted our beliefs about the effect size of ego depletion toward zero. Specifically, the posterior median was 0.087 with a central 95% credible interval ranging from 0.023 to 0.152 (Figure S3).

Figure S2. *Bayesian Forest Plot of Performance Outcome by Laboratory: Full Sample.* The values listed under BF_{+0} indicate relative support for the depletion hypothesis versus a hypothesis that there is no effect. Diamonds indicate overall effect sizes from meta-analytic models using fixed-effects, random-effects, and one that combined both approaches.

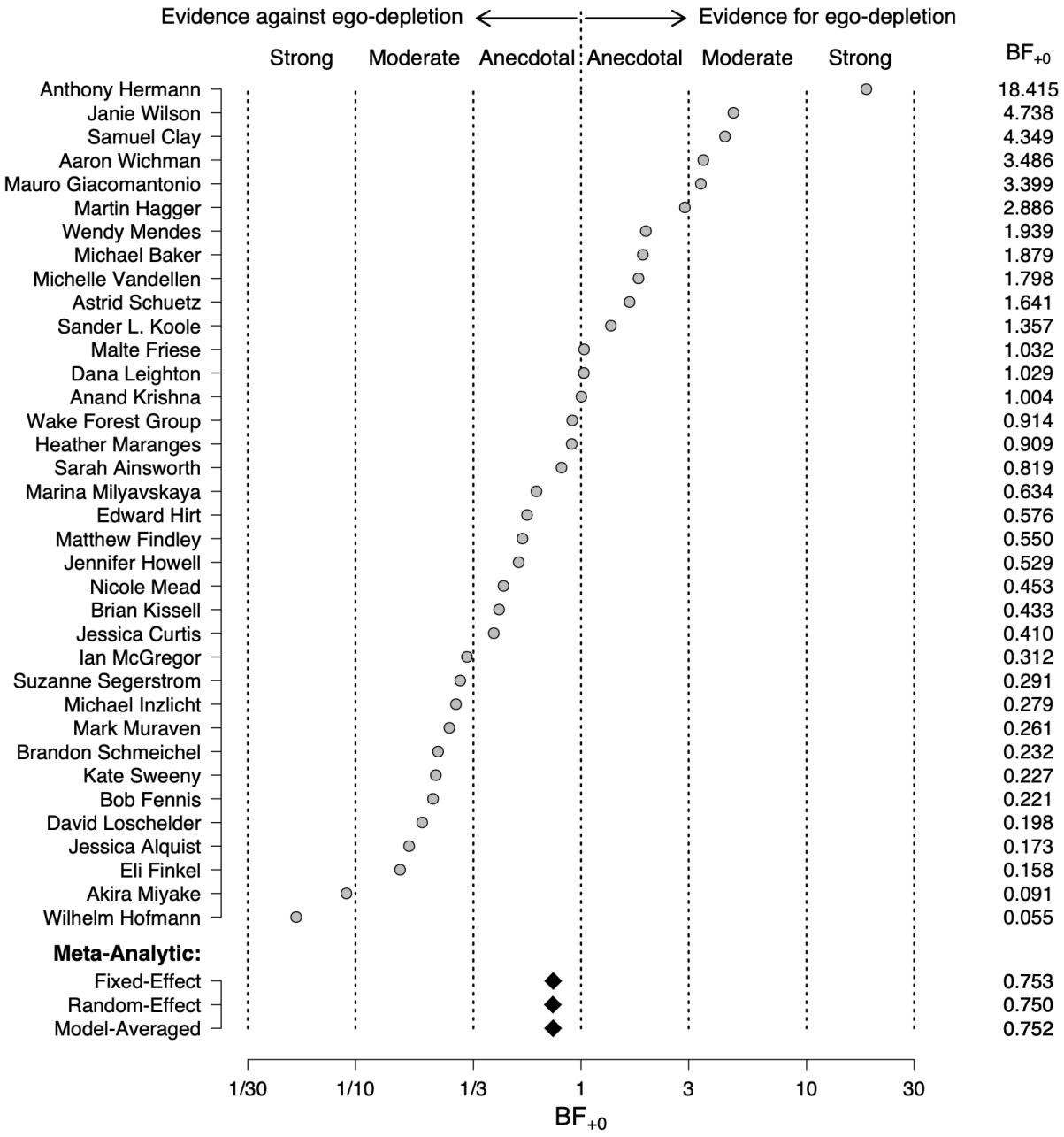
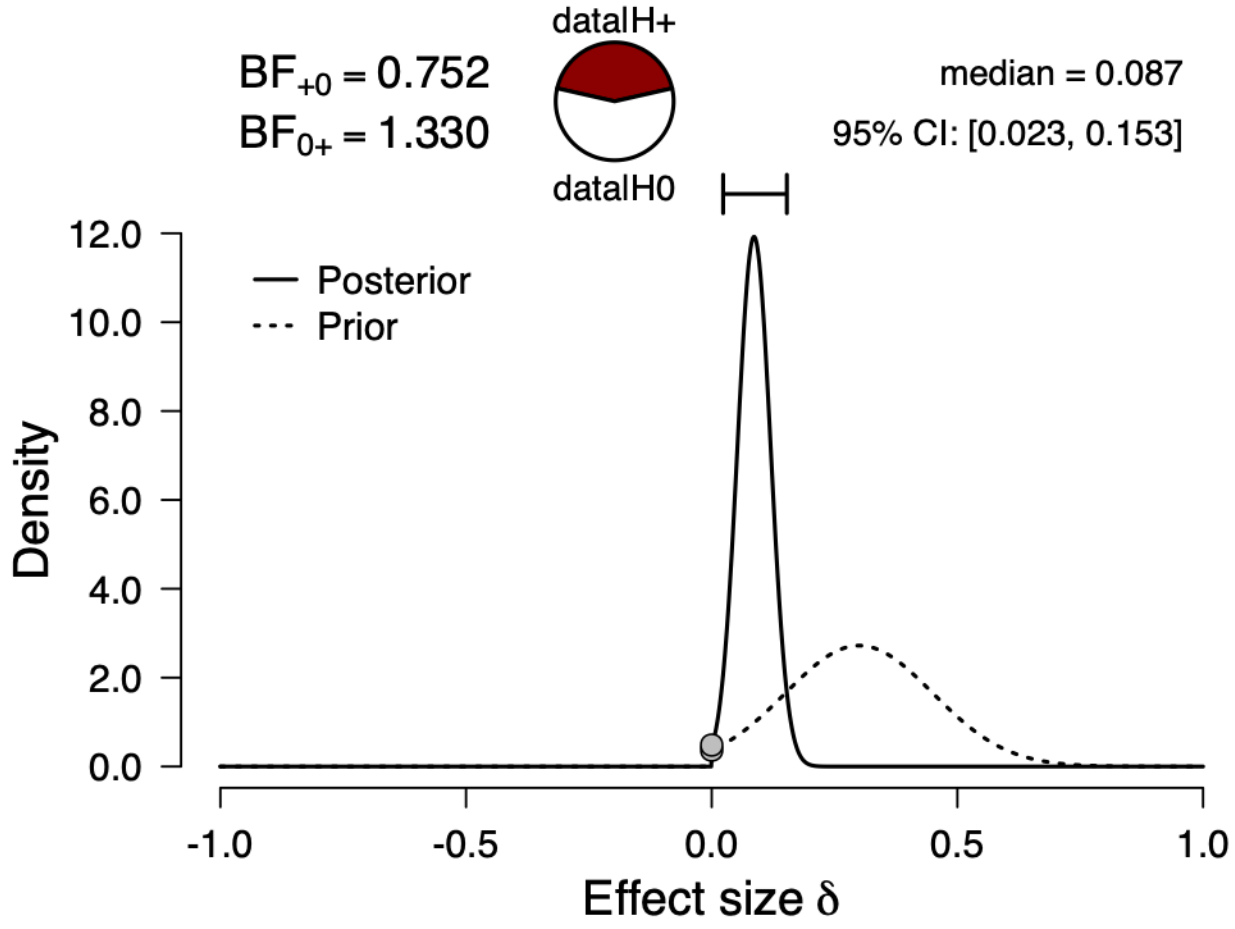


Figure S3. *Exploratory Tests of Model-Averaged Meta-Analytic Effect Size Posterior and Bayes Factor: Full Sample*. The dotted line indicates the informed prior effect size distribution and the solid line indicates the model-averaged meta-analytic posterior effect size distribution. Roughly-speaking, the peak of the shape indicates the likelihood of the effect size and its width indicates variance.



Moderators of the Depletion Effect

Protocol type. The main article reports confirmatory meta-analytical tests on the reduced sample (after preregistered exclusions; see Table 4). Here, we supplant those with parallel, exploratory results on the full sample and multi-level regressions on both samples.

Full sample: We examined the two components of the figure tracing task separately, the number of sheets participants used (as an indicator of attempts) and time spent working on the task (in seconds). Examining the two components separately, the effect of depletion condition on number of attempts was not statistically significant (Table S1; unstandardized descriptives: non-depletion condition *M* = 19.71, *SD* = 10.05; depletion condition *M* = 19.09, *SD* = 10.21).

There was a significant effect of depletion condition on duration in the full sample (unstandardized descriptives: non-depletion condition *M* = 988.87s, *SD* = 283.95; depletion condition *M* = 960.03s, *SD* = 298.52). These exploratory analyses showed that participants in the depletion condition gave up on the figure tracing task around 28s sooner than participants in the non-depletion condition (Table S1).

For the combined measure of figure tracing duration and attempts in the E-task protocol, there was a statistically significant effect in the full sample, as judged by both the meta-analytic and multi-level regression approaches. Participants in the depletion condition had lower figure tracing scores than did participants in the non-depletion condition (Table S1).

We conducted meta-analytic and multi-level analyses within the writing task protocol, which used the Cognitive Estimation Test (CET) as the performance measure.

The results were non-significant (unstandardized descriptives: non-depletion condition *M* = 1.32, *SD* = 0.23; depletion condition *M* = 1.31, *SD* = 0.24; Table S1).

Reduced sample. Multi-level regression models analyzed the reduced sample's performance within each protocol. For overall figure tracing scores, the effect of condition was not significant, b = 0.17, 95% CI [-0.01, 0.34].

As mentioned, that score has two elements. Breaking them down, the effect of condition on number of attempts was not statistically significant (*b* = 0.06, 95% CI [-.05, 0.17]; unstandardized descriptives: non-depletion condition *M* = 19.87, *SD* = 9.92; depletion condition *M* = 19.36, *SD* = 10.41).

As in the full sample, the effect of depletion condition on duration was significant in the reduced sample (*b* = 0.11, 95% CI [.01, 0.21]; unstandardized descriptives: non-depletion condition *M* = 1012.20s, *SD* 266.30; depletion condition *M* = 985.10s, *SD* = 283.52).

A last set of exploratory analyses regarded the depletion manipulation's effect on CET performance. As in the full sample, the effect of depletion condition was non-significant in the reduced sample (*b* = 0.01, 95% CI [-.09, 0.12]; unstandardized descriptives: non-depletion condition *M* = 1.34, *SD* = .23; depletion condition *M* = 1.34, *SD* = .23).

States and traits. We also examined whether self-reported states captured by the manipulation check items (e.g., fatigue, effort) and individual difference measures (i.e., trait self-control; willpower beliefs; action orientation) acted as moderators of the depletion effect. Because self-reported traits and states are best modeled as individual-level data, multilevel regressions were used as opposed to meta-analytic analyses (Table S2).

The only significant moderator was the fatigue index, which was evident in both the reduced and full samples. The depletion effect was larger for participants who reported being more fatigued by the manipulation task (Figures S4 and S5).

For the reduced sample (after exclusions), simple-slope analyses revealed that within the range of the data, the depletion effect was significant in a region from a standardized score of 0.15 on the fatigue index to the sample maximum of 2.37 (Figure S4). The magnitude of the depletion effect was $b = 0.23$, $SE = 0.07$, $p = .001$, at the 75th percentile of fatigue (0.84). For the full sample, the magnitude of the depletion effect at the 75th percentile of fatigue (0.84) was $b = 0.21$, $SE = 0.06$, $p < .001$.

Table S2. *Potential Moderators of the Depletion Effect: Frequentist Multi-Level Models*

				Intercept		Depletion manipulation		Moderator		Interaction	
Moderator variable	**Moderator type**	**Sample**	***N***	***b***	**CI**	***b***	**CI**	***b***	**CI**	***b***	**CI**
Protocol[a]	Study design	Reduced	2461	0.08	[-0.07, 0.23]	0.09	[-0.01, 0.19]	0.02	[-0.28, 0.32]	-0.16	[-0.36, 0.05]
		Full	3524	-0.04	[-0.19, 0.10]	0.11*	[0.02, 0.20]	0.06	[-0.22, 0.34]	-0.14	[-0.32, 0.04]
Effort index	Manipulation check	Reduced	2461	0.09	[-0.07, 0.24]	0.03	[-0.09, 0.16]	-0.02	[-0.11, 0.07]	-0.07	[-0.22, 0.09]
		Full	3523	-0.03	[-0.17, 0.11]	0.04	[-0.07, 0.15]	-0.03	[-0.11, 0.05]	-0.07	[-0.21, 0.06]
Fatigue index	Manipulation check	Reduced	2461	0.10	[-0.05, 0.24]	0.08	[-0.02, 0.18]	-0.15***	[-0.23, -0.07]	0.18**	[0.07, 0.29]
		Full	3523	-0.03	[-0.16, 0.11]	0.09	[-0.00, 0.18]	-0.15***	[-0.21, -0.08]	0.14**	[0.05, 0.24]
Frustration	Manipulation check	Reduced	2459	0.12	[-0.03, 0.27]	0.02	[-0.13, 0.10]	-0.11**	[-0.18, -0.03]	0.06	[-0.07, 0.19]
		Full	3521	-0.00	[-0.14, 0.14]	0.03	[-0.07, 0.13]	-0.10**	[-0.17, -0.04]	0.03	[-0.08, 0.14]
Action Orientation	Individual difference	Reduced	2356	0.08	[-0.07, 0.24]	0.08	[-0.02, 0.19]	-0.12	[-0.43, 0.20]	-0.07	[-0.51, 0.37]
		Full	3395	-0.04	[-0.18, 0.11]	0.11*	[0.02, 0.20]	-0.17	[-0.44, 0.10]	-0.03	[-0.42, 0.35]
Implicit Willpower Theory	Individual difference	Reduced	2341	0.05	[-0.10, 0.20]	0.09	[-0.01, 0.19]	0.01	[-0.07, 0.10]	0.02	[-0.09, 0.14]
		Full	3315	-0.05	[-0.19, 0.10]	0.09	[-0.00, 0.18]	0.02	[-0.06, 0.09]	0.04	[-0.06, 0.14]
Trait Self-Control	Individual difference	Reduced	2444	0.07	[-0.08, 0.22]	0.10	[-0.00, 0.20]	0.00	[-0.12, 0.12]	0.01	[-0.15, 0.17]
		Full	3490	-0.05	[-0.19, 0.09]	0.12**	[0.03, 0.21]	-0.01	[-0.11, 0.10]	0.03	[-0.11, 0.17]

Note: These tests are exploratory. Sample sizes vary due to missing data. Condition coded such that 0 = non-depletion, 1 = depletion condition. Results are raw beta weights (*b*) from random-effects multi-level mixed models; CI indicates 95% confidence intervals. Participants' data were nested within lab with random intercepts for labs and separate regression models were used for each moderator. Individual differences scores were mean-centered. [a] Contrast-coded, -.5 = E-task, .5 = Writing task. * $p < .05$

Figure S4. *Exploratory Test of Moderation of Task Performance by Depletion Condition and Self-Reported Fatigue: Reduced Sample*. The figure represents the interaction of depletion condition x fatigue scores on task performance with 95% confidence bands. Task performance was standardized and ranged from -5.54 to 7.05 (only the region from -1 to 1 is displayed). The fatigue index is an average of standardized ratings of fatigue and tiredness.

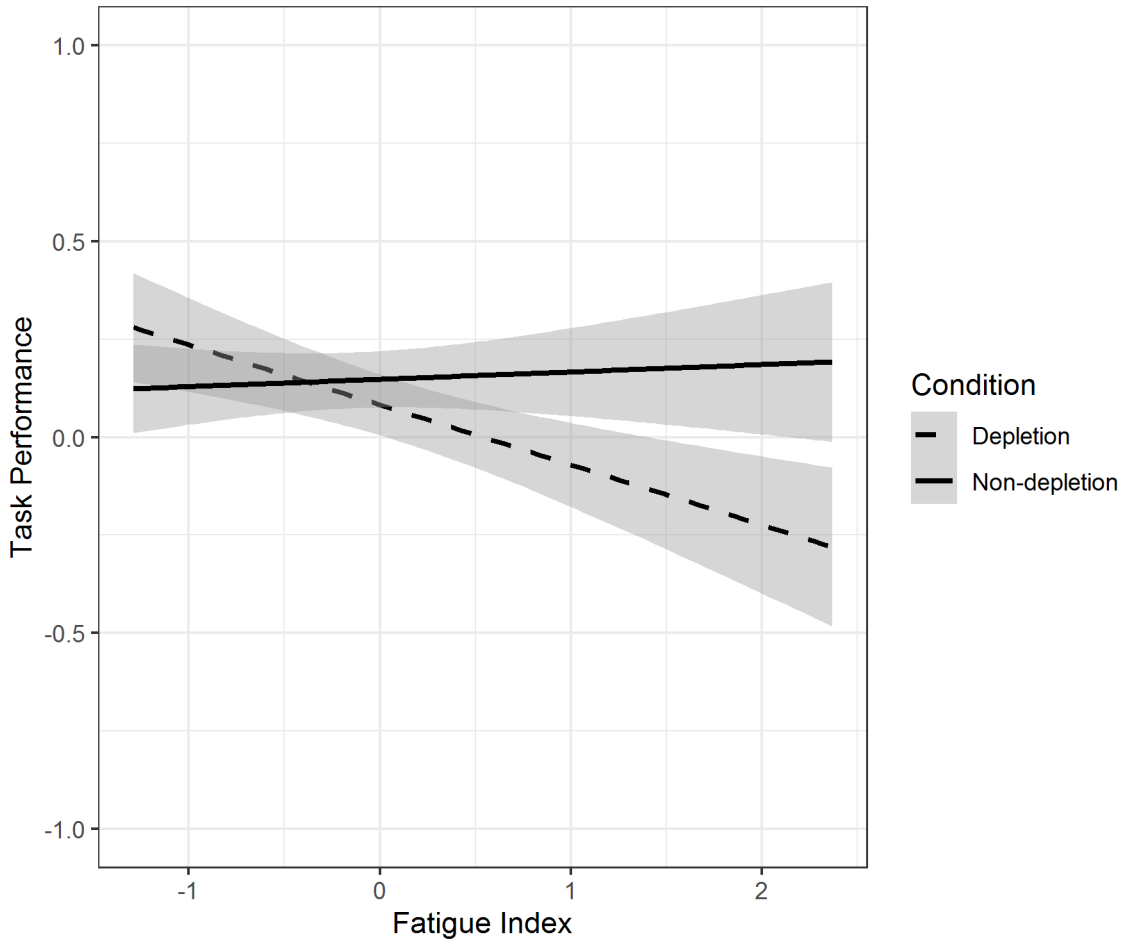
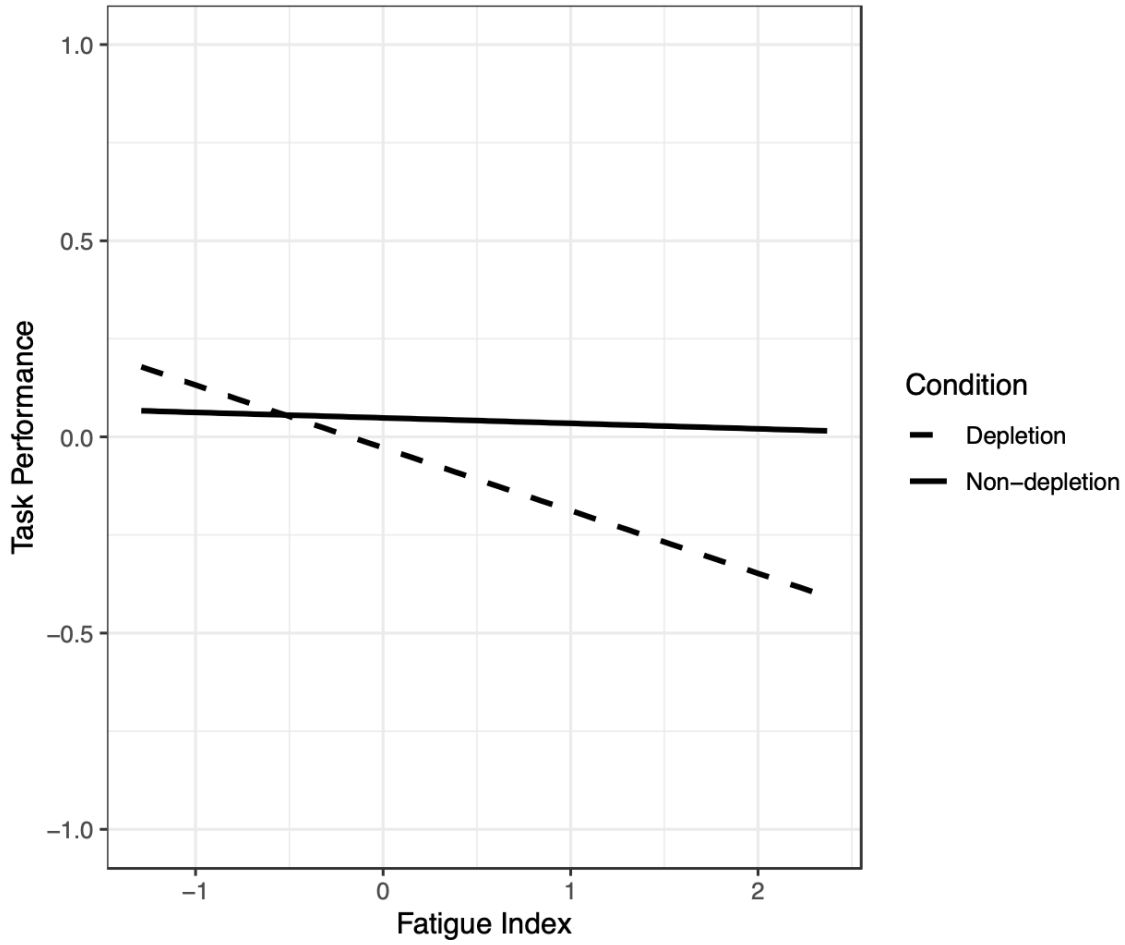


Figure S5. *Exploratory Test of Moderation of Task Performance by Depletion Condition and Self-Reported Fatigue: Full Sample*. The figure represents the interaction of depletion condition x fatigue scores on task performance with 95% confidence bands. Task performance was standardized and ranged from -5.54 to 7.05 (only the region from -1 to 1 is displayed). The fatigue index is an average of standardized ratings of fatigue and tiredness.



Secondary moderator analyses. We tested whether the depletion effect was moderated by: the number of depletion studies published by the principal investigator (PI) through 2016 (as counted independently by KV and BS), the number of total publications by the PI through 2016 (as counted by KV and BS), experimenter behavior (as rated by two independent coders of the videos submitted by each laboratory), and laboratory location (North American countries versus other countries). The latter moderator was chosen because many published depletion studies were conducted in North America so it was plausible that location might make a difference in the outcome.

The only significant moderator in these analyses was the role of experimenter behavior in the full sample (Table S3; Figure S6). Coding and composite score details are reported in the Additional Sample and Methodological Details section below. Experimenter behavior was not a significant moderator in the meta-analytic results on the full sample nor in meta-analytic or multi-level regression results on the reduced sample.

Exploratory multi-level regression analyses using the full sample showed an additional interaction of depletion condition and codings of experimenter behavior on task performance, b = -0.25, 95% CI [-0.45, -0.05]. The main effect of condition was significant in this model, b = 0.11, 95% CI [0.02, 0.20], and so was the main effect of experimenter behavior scores, b = 0.22, 95% CI [0.00, 0.43]. Simple slopes analyses of the interaction showed that experimenter behavior had no effect on performance in the non-depletion condition, b = -0.03, SE = 0.11, p = .776, but in the depletion condition, performance was worse when experimenters' behavior was rated lower, b = 0.22, SE = 0.11, p = .046. For experimenter behavior scores at the sample median (0.16) or above

(that is, experimenters who were at least moderately professional, at ease, and stuck to the script), there was no depletion effect, b = 0.06, SE = 0.05, p = .183. Below-average experimenter behavior scores were however related to the magnitude of the depletion effect, b = 0.18, SE = 0.06, p = .001, at the 25th percentile of experimenter behavior scores (-0.29).

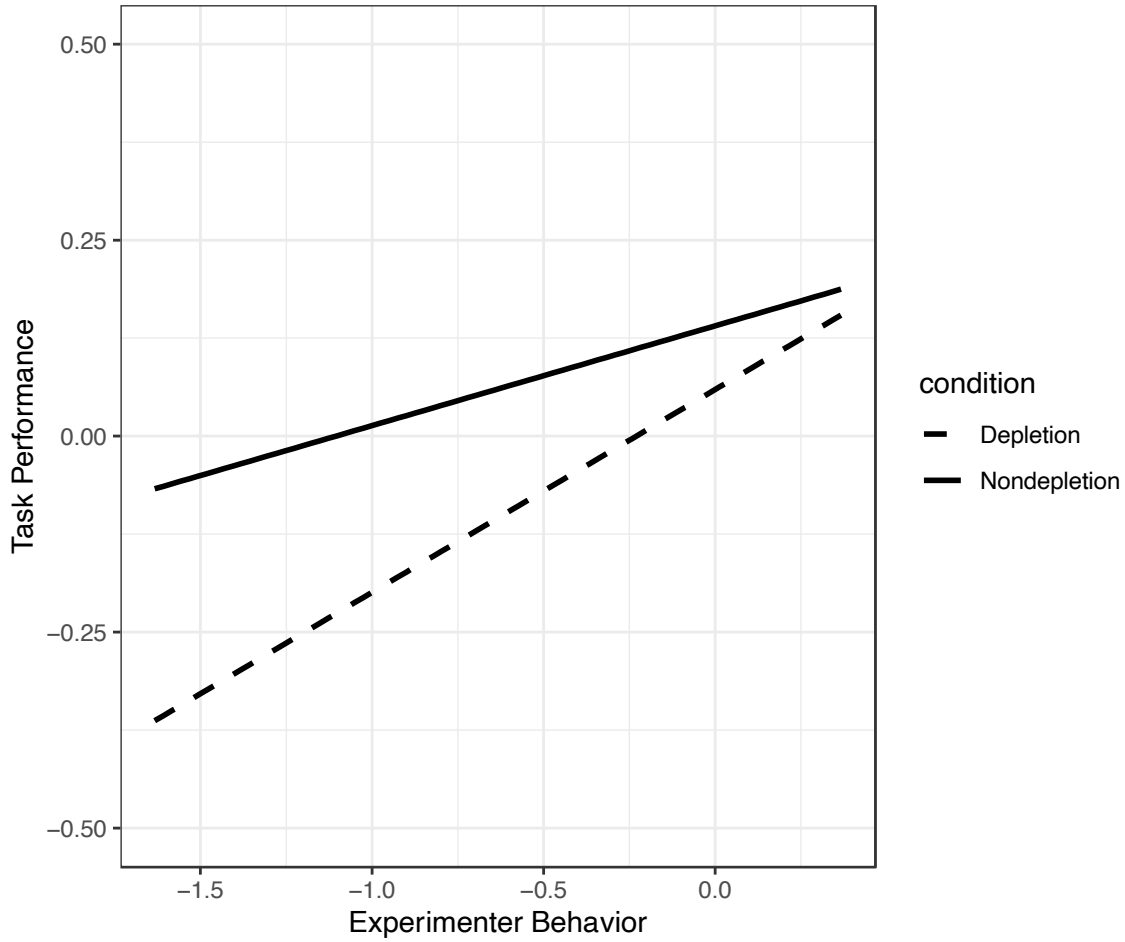
We preregistered our intention to test moderation of the depletion effect by experimenters' awareness of the depletion hypothesis or whether investigators had a Ph.D. We did not conduct these analyses because we did not solicit experimenters' knowledge of the depletion hypothesis prior to some laboratories initiating data collection and because there was very little variance in highest degree obtained. We also preregistered that we would test whether exclusion of participants based on the dependent measure differed as a function of depletion condition. The test, however, turned out to be inapplicable because the exclusion criteria were not set up to enable it.

Table S3. *Potential Depletion Effect Moderators: Exploratory Frequentist Multi-Level Models*

			Intercept		Depletion manipulation		Moderator		Interaction	
Moderator variable	**Sample**	***N***	***b***	**CI**	***b***	**CI**	***b***	**CI**	***b***	**CI**
Experimenter behavior	Reduced	2396	0.08	[-0.07, 0.23]	0.09	[-0.01, 0.19]	0.24*	[0.00, 0.48]	-0.19	[-0.42, 0.04]
	Full	3441	-0.04	[-0.18, 0.10]	0.11*	[0.02, 0.20]	0.22*	[0.00, 0.43]	-0.25*	[-0.45, -0.05]
Depletion studies count	Reduced	2461	0.08	[-0.09, 0.25]	0.13*	[0.01, 0.24]	-0.00	[-0.02, 0.01]	-0.01	[-0.01, 0.00]
	Full	3524	-0.07	[-0.23, 0.09]	0.15*	[0.05, 0.25]	0.00	[-0.01, 0.02]	-0.01	[-0.01, 0.00]
Publication count	Reduced	2461	0.08	[-0.07, 0.22]	0.09	[-0.01, 0.19]	0.00	[-0.00, 0.01]	-0.00*	[-0.01, -0.00]
	Full	3524	-0.05	[-0.19, 0.09]	0.11*	[0.02, 0.20]	0.00	[-0.00, 0.00]	-0.00	[-0.00, 0.00]
Lab location[a]	Reduced	2461	0.12	[-0.16, 0.40]	0.06	[-0.14, 0.25]	-0.06	[-0.39, 0.27]	0.05	[-0.18, 0.27]
	Full	3524	-0.05	[-0.31, 0.22]	0.08	[-0.08, 0.25]	0.00	[-0.31, 0.31]	0.04	[-0.16, 0.24]

Note: These tests are exploratory. All moderators are at the level of the lab. Sample sizes depart slightly from total sample sizes due to missing data. Condition coded such that 0 = non-depletion, 1 = depletion condition. Participants' data were nested within lab with random intercepts for labs and separate regression models were used for each moderator. Experimenter behavior scores were mean-centered. Depletion studies count was the number of published depletion studies by each Primary Investigator. Results are raw beta weights (*b*) from random-effects multi-level mixed models; CI indicates 95% confidence intervals. [a] Dummy-coded, 0 = Outside North America, 1 = North America. * $p<.05$

Figure S6. *Exploratory Test of Moderation of Task Performance by Depletion Condition and Experimenter Behavior: Reduced Sample*. The figure represents the interaction of depletion condition x judges' ratings of experimenter behavior (from video recordings) on task performance with 95% confidence bands. Task performance was standardized and ranged from -5.54 to 7.05 (only the region from -1 to 1 is displayed). Experimenter behavior is a composite of judges' ratings of experimenter professionalism, ease/comfort, and, for laboratories that conducted the study in English, adherence to the script.



Full sample manipulation checks

We conducted the same meta-analytic tests reported in the main text on the full sample of participants (i.e., no exclusions). Using the index of effort and difficulty ratings, the manipulation worked as intended (Table S4). We tested whether effort ratings differed by protocol, coded such that the intercept (d = 1.69, 95% [1.59, 1.79], I^2 = 36.09%) represents the average effect across both protocols (-.5 = E-task; .5 = Writing task). As in the confirmatory reduced sample tests, protocol was an unexpected moderator of manipulation check scores, b = 2.46, 95% CI [2.26, 2.67]. Although the depletion task was more difficult and effortful than the non-depletion task in both protocols, the difference was substantially larger in the writing task protocol compared to the E-task protocol.

We analyzed other task self-reports in a similar manner. The fatigue index revealed higher scores in the depletion condition than in the non-depletion condition. Similarly, reports of frustration were higher among depletion compared to non-depletion participants. Scores on the motivation index again did not differ by condition (Tables S4 and S5).

Exploratory tests of whether the manipulation check reports were moderated by protocol revealed some unanticipated patterns (Table S5). Reports on the effort index were moderated by protocol for both samples. For the reduced sample, that test was preregistered as it comprised the primary check of the manipulation (Table 3 in the main article). Protocol moderated scores on the fatigue index, such that in the writing task protocol, participants in the depletion condition reported being more fatigued than participants in the non-depletion condition, whereas in the E-task protocol, participants

in the non-depletion condition reported being more fatigued than participants in the depletion condition. The latter pattern runs contrary to expectations and the published literature (e.g., Baumeister, Bratslavsky, Muraven, & Tice, 1998; Legault, Green-Demers, & Eadie, 2009). Scores on the motivation index also were moderated by protocol. In the writing task protocol, participants in the depletion condition reported being more motivated than did participants in the non-depletion condition, which is another unexpected pattern. Motivation reports did not differ by condition in the E-task protocol. Frustration reports were not moderated by protocol.

We hesitate to speculate about the unexpected patterns for the fatigue and motivation indices, but there may be a few implications. An examination of the conditional means on the fatigue index suggests that the non-depletion task in the E-task protocol was not the clean, neutral exercise we assumed it would be. The motivation index difference, with participants in the controlled writing (versus free writing) condition reporting more motivation, is not consistent with any existing models of the ego depletion effect. The unexpected results from exploratory analyses of the manipulation checks would need to be replicated in future research to bolster confidence in them.

Table S4. *Manipulation Checks: Descriptive Statistics and Exploratory Frequentist Meta-Analytic Tests of Experimental Condition, Full Sample*

Variable	*M* (*SD*)	FE Average	CI	RE Average	CI	I^2
Effort index	3.56 (1.74)	1.21**	[1.14,1.29]	1.59**	[1.13, 2.03]	96.88%
Frustration	2.98 (1.94)	0.88**	[0.80, 0.95]	1.01**	[0.70, 1.33]	94.60%
Fatigue index	3.12 (1.56)	0.26*	[0.20, 0.33]	0.27*	[0.12, 0.42]	80.17%
Motivation index	5.23 (1.25)	0.04	[-0.03, 0.11]	0.04	[-0.04, 0.11]	20.84%

Note: *N* = 3528, with the exception that frustration ratings were missing for two participants. Sample size departs slightly from total sample size due to missing data. Condition coded such that 0 = non-depletion, 1 = depletion condition. Higher numbers indicate that participants in the depletion condition reported stronger feelings than participants in the non-depletion condition. All tests were exploratory. *M*s and *SD*s are from unstandardized scales ranging from 1 (*not at all*) to 7 (*very*). FE indicates fixed-effects models; RE indicates random-effects models. CI indicates 95% confidence intervals. * $p<.05$; ** $p<.01$

Table S5. *Manipulation Checks: Descriptive Statistics and Exploratory Frequentist Meta-Analytic Tests of Experimental Condition by Protocol, Full Sample*

			M (*SD*)				RE Average		
Variable	**Sample**	**Task**	**Depletion**	**Non-Depletion**	***k***	***N***	***d***	**CI**	**I^2 %**
Effort Index	Reduced	Writing Task	5.81 (1.09)	2.48 (1.03)	16	1246	3.09***	[2.87, 3.30]	39.29
		E-Task	3.27 (1.29)	2.71 (1.18)	20	1217	0.46***	[0.34, 0.57]	0
	Full	Writing Task	5.77 (1.13)	2.52 (1.05)	16	1679	2.98***	[2.71, 3.25]	72.48
		E-Task	3.33 (1.34)	2.74 (1.23)	20	1849	0.45***	[0.36, 0.55]	0
Frustration	Reduced	Writing Task	5.05 (1.65)	1.77 (1.24)	16	1246	2.26***	[2.07, 2.46]	45.04
		E-Task	2.62 (1.55)	2.34 (1.48)	20	1215	0.19**	[0.06, 0.32]	22.08
	Full	Writing Task	4.98 (1.74)	1.89 (1.34)	16	1679	2.01***	[1.84, 2.19]	51.81
		E-Task	2.74 (1.62)	2.42 (1.52)	20	1847	0.19***	[0.10, 0.29]	0
Fatigue Index	Reduced	Writing Task	3.24 (1.59)	2.29 (1.31)	16	1246	0.67***	[0.52, 0.83]	43.08
		E-Task	3.33 (1.47)	3.53 (1.50)	20	1217	-0.15*	[-0.29, -0.01]	30.61
	Full	Writing Task	3.30 (1.61)	2.29 (1.33)	16	1679	0.70***	[0.59, 0.80]	14.61
		E-Task	3.35 (1.51)	3.47 (1.49)	20	1849	-0.10	[-0.20, 0.00]	18.00
Motivation Index	Reduced	Writing Task	4.87 (1.19)	4.62 (1.22)	16	1246	0.19**	[0.07, 0.31]	12.52
		E-Task	5.61 (1.10)	5.70 (1.06)	20	1217	-0.10	[-0.22, 0.01]	1.85
	Full	Writing Task	4.85 (1.22)	4.65 (1.22)	16	1679	0.14**	[0.04, 0.24]	8.10
		E-Task	5.64 (1.11)	5.69 (1.11)	20	1849	-0.06	[-0.15, 0.04]	8.34

Note: Condition coded such that 0 = non-depletion, 1 = depletion condition. Higher numbers indicate that participants in the depletion condition reported stronger feelings than participants in the non-depletion condition. All tests were exploratory. *M*s and *SD*s are from unstandardized scales ranging from 1 (*not at all*) to 7 (*very*). RE indicates random-effects models. CI indicates 95% confidence intervals.
* $p<.05$; ** $p<.01$; *** $p<.001$

Additional Sample and Methodological Details

Recruitment

The lead author (KV) announced the intention to conduct this replication on behavioral science listservs. She also sent personal emails to prominent scholars who have published on ego depletion, including to scholars who have been publicly critical of depletion. Forty laboratories indicated commitment to participating in the project. Six dropped out before initiating or completing data collection and two additional laboratories joined before the end of the data collection period.

Materials and procedures

Participating laboratories received a script for how to conduct the experiment, complete with the wording they should use and the arrangement of the laboratory. When necessary, members of non-English-speaking laboratories translated the script and experimental materials into the language in which the study would be conducted. Additionally, KV created video samples of how to conduct each protocol and shared them with participating labs. Via Skype, KV or BS communicated with laboratories to answer questions and provide additional information. Last, laboratories in both protocols were instructed to have experimenters leave the room while participants performed the study's tasks (independent variable task, dependent variable task, manipulation check ratings, individual difference measures, demographics, and post-experimental questionnaires).

E-task protocol. The instructions for both pages of this task were in the laboratory's native language whereas the E-task text was in English for all participants (even if the laboratory's native language was not English).

Participating laboratories reported the number of errors participants made on the last full paragraph participants completed of the manipulation task used in the E-task protocol. Crossing out an E that should have been skipped and skipping an E that should have been crossed out both counted as errors.

A figure-tracing task served as the dependent measure in this protocol. Experimenters surreptitiously recorded how long participants persisted at figure tracing and counted the number of figure sheets participants attempted to solve.

Story-writing protocol. Laboratories reported uses of forbidden letters (i.e., *a* and *n*) and simple omissions of forbidden letters (e.g., “the dog b_rked”) for each participant in the depletion condition of the story-writing protocol. Only depletion condition participants could have errors. Across both conditions, story word counts were reported.

The CET served as the dependent measure in this protocol. Experimenters timed the duration participants took to complete the CET.

We did not include one item from the published version of the CET, “How much does a telephone weigh?” The published scoring metric (see Bullard et al., 2004; Fein et al., 1998) does not correspond to the weight of contemporary telephones. Additionally, some items on the CET ask for imperial measurements (e.g., “How many sticks of spaghetti are there in a one pound package?”). For labs outside North America, those items were revised to indicate the metric system.

Responses to each item were converted to a common metric before final scoring of the CET. The CET was scored using published norms (Bullard et al., 2004; Fein et al., 1998). Answers within 25-75% of the normative range (i.e., good estimates) received 2

points. Answers outside the 25-75% range but within the 5-95% normative range received 1 point. Answers outside the normative range (i.e., extreme estimates) received 0 points. Participants occasionally gave answers with a tilde (e.g., ~1), which we treated as the numerical value (e.g., 1). Responses given as a range (e.g., 6 to 8) were treated as the median of the two values (e.g., 7).

We considered some answers invalid. Some items did not specify a unit of measurement (e.g., distance could be reported in inches, feet, miles, and so on), and participants were instructed to provide the unit of their response. If they did not provide a unit of measurement for a relevant item, the response was considered invalid. If participants did not report a numerical answer (e.g., “infinite”) or provided a nonsensical answer (e.g., “0.5 pounds” for an item asking for a number of spaghetti sticks), the response was considered invalid. Last, if participants skipped an item, it counted as invalid. The final CET score for each participant was an average calculated by summing item scores and dividing by the number of valid responses.

Videos of experimenters. All but two labs submitted recordings of experimenters conducting the study on a practice subject, although five lacked usable audio or video. A total of 65 videos were coded by two independent coders using scales from 1 (*not at all*) to 5 (*very much*) on professionalism (i.e., how competent, in charge, like a leader, and professional in appearance the experimenter behaved), r = 0.70, 95% CI [0.68, 0.73], κ = 0.63, M = 4.64, SD = 0.49), and ease/comfort (i.e., how warm, natural, comfortable, and not stiff or robotic the experimenter behaved), r = 0.53, 95% CI [0.50, 0.55], κ = 0.36, M = 4.56, SD = 0.56). For labs that conducted the study in English, videos (n = 49) also were coded for adherence to the script (r = 0.72, 95% CI

[0.69, 0.74], κ = 0.44, *M* = 4.61, *SD* = 0.65). Judges' ratings of professionalism, ease/comfort, and adherence to the script were averaged and then combined into a composite score of experimenter behavior. (The composite score for laboratories that did not conduct the study in English, and hence did not have ratings for script adherence, was comprised of professionalism and ease/comfort ratings.) Descriptive statistics for the video codings were based on the full sample of participants.

Exclusions

Following preregistered criteria, we excluded data from *n* = 1068 participants as follows. (Some participants failed multiple exclusion criteria.) The overall number of participants who were excluded was more than we expected, but by percentage of all participants the exclusion rate aligns closely with another multi-site depletion replication study. Hagger et al.'s (2016) multi-lab depletion replication paper reported an exclusion rate of 30.9% (*n* = 958 out of 3099 total participants). By comparison, our exclusion rate was 30.25% (1068 out of a total sample size of 3531).

The exclusion criteria can be broadly understood as belonging to four categories: 1) participants' performance errors or mistakes on the tasks (e.g., errors on the E-task, invalid responses on the CET), 2) participants' behavior (e.g., being disturbed, disruptive, or disrupted; using their phone in violation of instructions; knowing that the puzzles were unsolvable in the E-task protocol), 3) participant characteristics (being a non-native speaker of the language in which the study was run; being one of the experimenters' first three participants), and 4) other exclusions. Experimenters noted irregularities that occurred during the course of the study, and three independent coders determined whether each irregularity qualified as an exclusion. (For more information on

that process, see below under “Both protocols.”) Examples of issues determined to be disqualifying included noise from construction during the study, a repeat participant, missing the timing cue to stop a task, and experimenters being acquainted with participants. Counts of excluded participants based on each preregistered criterion are reported in Table 2 in the main article.

E-task protocol. We excluded data from participants who made more than 2.5 MAD (median absolute deviation) errors on the last full paragraph they completed on the E-crossing task (Leys, Ley, Klein, Bernard, & Licata, 2013). For page 1 of the task (the habit-forming portion), MAD calculations were done at the lab level. For page 2 of the task (the habit-breaking portion), MAD calculations were done within lab and separately by condition. We also excluded data from participants who expressed knowledge (prior to the debriefing) that the figures used in the figure-tracing task (the dependent measure in this protocol) were unsolvable. Table 2 in the main text displays exclusion counts.

Story writing protocol. We excluded data from participants who used 2.5 MAD or fewer words than other participants in their lab and in the same experimental condition, participants who used the restricted letters (*a* and *n*) more often than 2.5 MAD of the lab (this criterion applied only to the depletion condition), and participants who scored beyond 2.5 MAD of the lab mean on invalid responses on the CET (Table 2).

Both protocols. As preregistered, we excluded participants who were non-native speakers as indicated by matching the language(s) they reported speaking at home against the language in which the study was run, who were among the first three run by

each experimenter, who reported using their phone during the study, and who were reported by the experimenter to be belligerent, or distressed or distraught. Also as preregistered, we excluded data from participants who experienced a disruption during the experiment session or otherwise experienced an unanticipated deviation from the experimental procedures, as indicated by the experimenter (Table 2).

Further, we instructed experimenters to note other concerns that may warrant excluding the participant. That information was culled and sent to KV, BS, and Rebecca Schlegel, who independently coded whether the concerns merited exclusion of that participant's data. Coders were blind to all other data pertaining to the participant (e.g., condition, protocol, scores on the dependent measures). Exclusions occurred only when all three coders agreed that a participant should be excluded ("Other exclusions;" Table 2). In cases when two of the three coders thought a participant should be excluded, all coders conferred and came to a consensus.

Principal Investigators and Laboratory Members

*Ainsworth, Sarah E., Tallahassee Community College
Bunyi, Angelica, University of North Florida
*Fuglestad, Paul, University of North Florida
Hartsell, Bethany, University of North Florida

*Alquist, Jessica, L., Texas Tech University
Campbell, Collier, Texas Tech University
Price, Mindi, M., Texas Tech University
Stinnett, Alec, J., Texas Tech University
Tonnu, Karine, Texas Tech University

*Baker, Michael, D., East Carolina University
Walker, Jasmine, S., East Carolina University
White, Rachel, A., East Carolina University

*Clay, Samuel L., Brigham Young University-Idaho
Christensen, Weston, J., Brigham Young University-Idaho
Johnson, Hannah, L., Brigham Young University-Idaho
*Wiggins, Bradford, J., Brigham Young University-Idaho

*Curtis, Jessica, Arkansas State University
Johnson, Emily, Arkansas State University

*Hagger, Martin S., University of California, Merced and University of Jyväskylä
Chatzisarantis, Nikos, L. D., Curtin University
Lee, Nick, Curtin University
Meslot, Carine, Curtin University

*Hermann, Anthony, D., Bradley University
Hutton, Robert, D., Bradley University
Lee, Kelemen, T., Bradley University

*Hirt, Edward R., Indiana University
Eyink, Julie, R., Indiana University
Sherman, Janelle, Indiana University

*Howell, Jennifer L., University of California, Merced
Rockwell, Rachael, Ohio University
Sosa, Nicholas, Ohio University
Theodore, Dominic, Ohio University

*Fennis, Bob M., University of Groningen, the Netherlands
Gineikiene, Justina, AdCogito Institute for Advanced Behavioral Research, ISM University of Management and Economics

Hidding, Jasper, J., University of Groningen, the Netherlands
Joye, Yannick, ISM University of Management and Economics, Vilnius, Lithuania
Moeini-Jazani, Mehrad, University of Groningen, the Netherlands

*Findley, Matthew, B., Austin College
Mazara, Jr., Kennedy, Austin College

*Finkel, Eli, J., Northwestern University
Doğruol, Yasemin, Northwestern University

*Friese, Malte, Saarland University
Kaben, Jan Helge, Saarland University
Gieseler, Karolin, Saarland University

*Giacomantonio, Mauro, University “Sapienza” of Rome
Brizi, Ambra, University “Sapienza” of Rome
De Cristofaro, Valeria, University “Sapienza” of Rome
Salvati, Marco, University “Sapienza” of Rome

*Hofmann, Wilhelm, Ruhr University Bochum
Diel, Katharina, Ruhr University Bochum
Grande, Maria, University of Cologne
Stapels, Julia, University of Cologne

*Inzlicht, Michael, University of Toronto
Cau, Chuting, University of Toronto
Patel, Krishna, University of Toronto
Saunders, Blair, University of Dundee

*Kammrath, Lara, K., Wake Forest University
*Masicampo, E.J., Wake Forest University
*Petrocelli, John, V., Wake Forest University
*Scherer, Anne, Wake Forest University
*Song, Yu, Wake Forest University
*Waugh, Christian, E., Wake Forest University

*Kissell, Brian L., Central Michigan University
Gibson, Bryan, Central Michigan University

*Koole, Sander, L., VU Amsterdam
van Oldenbeuving, Yasmijn, VU Amsterdam
Weise, Feline, VU Amsterdam

*Krishna, Anand, University of Würzburg
Eder, Andreas B., University of Würzburg
Geraedts, Lea F., University of Würzburg

Russ, Isabella F., University of Würzburg

*Leighton, Dana, C. Texas A&M University, Texarkana

*Loschelder, David D., Leuphana University Lüneburg
Pollak, Katja, M., Leuphana University Lüneburg
Rath, Maximilian, Leuphana University Lüneburg

*Maranges, Heather, M., Florida State University
Ersoff, Mia, Florida State University
Gobes, Carina, M., Florida State University
Joyce, Sarah, M., Florida State University
Kelly, Caitlin, N., Florida State University
Vergara, Raiza, C., Florida State University

*McGregor, Ian, University of Waterloo
Sharpinskyi, Konstantyn, University of Waterloo
Wheeler, Craig, University of Waterloo

*Mead, Nicole L., Schulich School of Business, York University
Hodge, Josh, University of Melbourne
James, Lily, The University of the Arts London

*Mendes, Wendy B., University of California, San Francisco
del Rosario, Kareena, University of California, San Francisco
Nakahara, Erin, University of California, San Francisco

*Milyavskaya, Marina, Carleton University
Capaldi, Jonathan, Carleton University
Werner, Kaitlyn, M., Carleton University
Shaw, Meaghan, Carleton University

*Miyake, Akira, University of Colorado Boulder
Robertson, Jacob A., University of Colorado Boulder
Schmitt, Kristin N., University of Colorado Boulder

*Muraven, Mark, University at Albany
Donaldson, Tina L., University at Albany
McCarthy, Samantha, University at Albany
Serenka, Benjamin, University at Albany

*Schmeichel, Brandon J., Texas A&M University
Chambers, Heather, Texas A&M University
Finley, Anna, University of Wisconsin-Madison
Strawser, Hannah, R., Texas A&M University

*Schütz, Astrid, University of Bamberg

*Segerstrom, Suzanne C., University of Kentucky
Gloger, Elana, M., University of Kentucky
Garcia-Willingham, Natasha, E., University of Kentucky

*Sweeny, Kate, University of California, Riverside
Lam, Christine, University of California, Riverside
Spillane, Kaitlyn, University of California, Riverside
Falkenstein, Angelica, University of California, Riverside

*vanDellen, Michelle R., University of Georgia
Butschek, Grant, J., University of Georgia

*Wichman, Aaron L., Western Kentucky University
Ramsey, Haley, J., Western Kentucky University

*Wilson, Janie H., Georgia Southern University
Forgea, Victoria, Georgia Southern University

Note: Laboratories are listed under the name of the PI used in the tables and figures, followed by additional members. For ease of presentation, tables and figures refer to each laboratory using the name of a PI, although some groups had more than one PI. The Wake Forest laboratory considered all members to be PIs and therefore is listed by site.

* indicates laboratory PIs.

References

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.

Bullard, S. E., Fein, D., Gleeson, M. K., Tischer, N., Mapou, R. L., & Kaplan, E. (2004). The Biber cognitive estimation test. *Archives of Clinical Neuropsychology*, *19*, 835-846.

Fein, D., Gleeson, M. K., Bullard, S., Mapou, R., & Kaplan, E. (1998, February). *The Biber Cognitive Estimation Test*. Poster presented at the annual meeting of the International Neuropsychological Society, Honolulu, HI.

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

Legault, L., Green-Demers, I., & Eadie, A. L. (2009). When internalization leads to automatization: The role of self-determination in automatic stereotype suppression and implicit prejudice regulation. *Motivation and Emotion*, *33*, 10-24.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*, 764 –766. https://doi.org/10.1016/j.jesp.2013.03.013