Malte Friese, Julius Frankenbach, Veronika Job, and David D. Loschelder. Does Self-Control Training Improve Self-Control? A Meta-Analysis. Perspectives on Psychological Science. Prepublished August 28, 2017.
 DOI: 10.1177/1745691617697076.

Does self-control training improve self-control? A meta-analysis

Malte Friese
Saarland University

Julius Frankenbach
Saarland University

Veronika Job
University of Zurich

David D. Loschelder
Leuphana University of Lueneburg

Declaration of conflicting interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Author Note

Malte Friese, Department of Psychology, Saarland University; Julius Frankenbach, Department of Psychology, Saarland University; Veronika Job, Department of Psychology, University of Zurich; David D. Loschelder, Faculty of Business and Economics, Leuphana University of Lueneburg.

The first and second author contributed equally to this work. We thank Alexander Hart for his assistance in coding the studies included in the meta-analysis, Joanne Beames, Tom Denson, and Martin Hagger for their valuable input and comments, and Zachary Fisher as well as Elizabeth Tipton for their advice with the implementation of the robust variance estimation approach to analyze the data. We are indebted to all primary authors who generously provided us with additional information about their studies. All data, code, full documentation of our procedures, and additional analyses are available at https://osf.io/v7gxf/.

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

Abstract

Self-control is positively associated with a host of beneficial outcomes. Therefore, psychological interventions that reliably improve self-control are of great societal value. A prominent idea suggests that training self-control by repeatedly overriding dominant responses should lead to broad improvements in self-control over time. Here, we conducted a random-effects meta-analysis based on robust variance estimation of the published and unpublished literature on self-control training effects. Results based on 33 studies and 158 effect sizes revealed a small-to-medium effect of $g = 0.30$, CI_{95} [0.17, 0.42]. Moderator analyses found that training effects tended to be larger for (a) self-control stamina rather than strength, (b) studies with inactive compared to active control groups, (c) males than females, and (d) when proponents of the strength model of self-control were (co-)authors of a study. Bias-correction techniques suggested the presence of small-study effects and/or publication bias, and arrived at smaller effect size estimates (range: $g_{corrected} = .13$ to .24). The mechanisms underlying the effect are poorly understood. There is not enough evidence to conclude that the repeated control of dominant responses is the critical element driving training effects.

Word count: 178

Keywords: self-control training, intervention, meta-analysis, publication bias, robust variance estimation

Does self-control training improve self-control? A meta-analysis

Successful self-control is associated with a host of positive outcomes in life, including academic success, stable personal relationships, financial security, and good psychological and physical health. By contrast, poor self-control is associated with more aggression, substance use, and crime, among others (Duckworth & Seligman, 2005; Gottfredson & Hirschi, 1990; Tangney, Baumeister, & Boone, 2004). It is readily conceivable that how well people fare in these domains has not only important personal consequences, but also consequences for society at large. Research shows that self-control assessed very early in life predicts a variety of important life outcomes (Daly, Delaney, Egan, & Baumeister, 2015; Moffitt et al., 2011). These findings seem to suggest that self-control is a stable trait being shaped early in life. However, other research perspectives highlight the possibility of self-control change by targeted interventions (e.g., Piquero, Jennings, Farrington, Diamond, & Gonzalez, 2016). Over the past 15 years, researchers have designed controlled psychological interventions that tested the effect of self-control training on self-control success across diverse domains (Berkman, 2016). Given the importance of self-control in various life domains, there is a tremendous demand for such interventions that promise to reliably, appreciably, and enduringly improve self-control. The present article provides a meta-analysis of this self-control training literature.

What self-control is and why it should (not) be possible to improve it

One prominent conceptualization defines self-control as the "ability to override or change one's inner responses, as well as to interrupt undesired behavioral tendencies (such as impulses) and refrain from acting on them" (Tangney et al., 2004, p. 274). In line with this definition, the exertion of self-control is typically seen as deliberate, conscious, and effortful.

The main theoretical rationale why training self-control should be beneficial comes from the strength model of self-control (Baumeister & Vohs, 2016b; Baumeister, Vohs, &

Tice, 2007). This influential model proposes that all self-control efforts draw on a general capacity. This capacity is used and depleted regardless in which domain a person exerts self-control (e.g., attention control, control of food intake, control of emotional expression). Because of its generality, improvements in the general self-control capacity should benefit all kinds of self-control behavior across various domains.

The strength model posits that the capacity to exert self-control works akin to a muscle. This assertion has two important implications: First, exerting self-control will lead to temporary exhaustion and make subsequent self-control failure more likely (ego depletion).[1] Second, repeated practice will strengthen the self-control muscle (training hypothesis). This will result in either a general increase in absolute muscle strength (i.e., improved self-control *strength*) and/or increased resistance to fatigue when confronted with demands (i.e., improved self-control *stamina*). Both increases in strength or stamina should benefit self-control in a broad range of domains in the laboratory and in everyday life.

From the perspective of the strength model, the crucial aspect of a training regimen lies in the repeated overriding of dominant responses. In typical self-control training studies that are examined in the present meta-analysis, participants are asked to complete everyday activities with the non-dominant hand such as brushing teeth or using the computer mouse (Miles et al., 2016), to refrain from using highly prevalent slang words (Finkel, DeWall, Slotter, Oaten, & Foshee, 2009), or to work on computerized tasks requiring the control of dominant responses (Cranwell et al., 2014). After the training (typically two weeks long), laboratory or everyday-life indicators of self-control strength and stamina are compared to a control group. Training effects have been investigated on outcome variables such as success

[1] Recently, there has been substantial debate about the magnitude of the ego depletion effect (Baumeister & Vohs, 2016a; Carter, Kofler, Forster, & McCullough, 2015; Hagger et al., 2016; Inzlicht, Gervais, & Berkman, 2015). Details of this debate are beyond the scope of the present meta-analysis that is primarily concerned with the second implication of the muscle analogy, the trainability of self-control.

in quitting smoking (Muraven, 2010b), laboratory aggression (Denson, Capper, Oaten, Friese, & Schofield, 2011), or physical persistence (Cranwell et al., 2014).

The hypothesis that training self-control leads to broad improvements in self-control across domains is both intriguing and risky: It is *intriguing* because the trainability of self-control has implications for many subfields of psychology and is of high practical importance. Among other benefits, it would open the possibility of helping people deal with self-control problems in one domain by practicing self-control in a completely different domain. For instance, consider an obese person having gone through countless unsuccessful diets, still wishing to lose weight. At this point, any new intervention directly concerned with restraining eating behavior may be difficult, because dieting is closely associated with frustration and feelings of personal failure. The self-control training hypothesis is intriguing in that it suggests this person could succeed at dieting by practicing self-control in unrelated and emotionally uncharged activities.

The self-control training hypothesis is a *risky* hypothesis because other literatures on training psychological capabilities are not very encouraging concerning appreciable and broad benefits in people's lives. Consider the literature on cognitive training of executive functions such as working memory capacity or task shifting (Miyake & Friedman, 2012). This literature shows that the transfer of improvements in the specific training tasks to other tasks measuring the same construct (i.e., from one working memory task to the other) is sometimes found (near transfer). By contrast, transfer rarely emerges to related constructs (i.e., from working memory to task-shifting) or behaviors that should benefit from improving the focal construct (far transfer; Melby-Lervåg & Hulme, 2013; Melby-Lervåg, Redick, & Hulme, 2016; Owen et al., 2010; Shipstead, Redick, & Engle, 2012). The empirical studies that have been conducted to date to test the self-control training hypothesis have exclusively focused on far-transfer—training took place in one domain (e.g., controlling speech and/or

posture) and dependent variables were collected in different domains (e.g., persistence, aggression).

Within the self-control literature, related but distinct conceptualizations of self-control stress the importance of learning essential self-control skills early in life (Heckman, 2006; Mischel et al., 2011; Moffitt et al., 2011). For example, preschoolers can learn to conceive desired objects as less tempting by focusing on their nonconsummatory features (Mischel & Baker, 1975). Recent meta-analytic evidence suggests that teaching such self-control skills is effective in children and adolescents to improve self-control ($g = 0.32$) and to reduce delinquency ($g = 0.27$; Piquero et al., 2016). The self-control training interventions reviewed in the present meta-analysis focus on repeatedly overriding dominant responses without teaching strategies how to do so. This approach might be less effective to appreciably and enduringly improve self-control.

Previous meta-analyses

Two peer-reviewed meta-analyses have previously summarized evidence relating to the self-control training hypothesis. The first meta-analysis included a total of nine published studies and revealed a large average effect of $d^+ = 1.07$ (Hagger, Wood, Stiff, & Chatzisarantis, 2010). Among these nine studies were three studies with exceptionally large effects sizes up to $d^+ > 8$ (sic!) and unclear methodology (Oaten & Cheng, 2006a, 2006b, 2007), leading to a very wide 95% confidence interval for the estimated average effect size [0.10, 2.03]. A more recent meta-analysis excluded these three studies, and included a total of ten published studies (Inzlicht & Berkman, 2015). Inzlicht and Berkman used the recently introduced *p*-curve method (Simonsohn, Nelson, & Simmons, 2014) to compute two estimates of the meta-analytic self-control training effect size—one based on the first dependent variable reported for a given study, the other based on the last dependent variable reported. All other effects were discarded. The first estimate was $d = 0.17$, CI_{95} [-0.07, 0.41],

a small effect not significantly different from zero. The second estimate was $d = 0.62$, CI_{95} [0.13, 1.11], a stronger, but also more volatile effect size.[2]

The present meta-analysis

The present meta-analysis aims to deliver a comprehensive summary of the published and unpublished evidence, and to considerably extend previous work. In particular, we pursued three goals: First, we aimed at estimating the average self-control training effect based on the most comprehensive data base possible. With 33 studies (23 published, 10 unpublished) we included more than three times as many studies than the Hagger et al. (2010) and the Inzlicht and Berkman (2015) meta-analyses. In addition, we based our estimates on *all* reported dependent variables, an issue of importance given that many of the original studies reported several dependent variables. In such cases, basing effect size estimates solely on the first and/or last reported effect (Inzlicht & Berkman, 2015) inevitably implies a loss of valuable information.

Second, we sought to conduct moderator analyses to elucidate boundary conditions of the self-control training effect. Moderator analyses can be crucially informative for both theory building and for applied purposes when designing self-control training procedures.

Finally, we sought to investigate the existence of small-study effects and publication bias. Publication bias accrues when studies with a statistically significant result are more likely to be published than studies with a null result. Because publishing almost exclusively significant results is how the field worked for many years (Bakker, van Dijk, & Wicherts,

[2] Two further recent meta-analyses examined effects of computerized inhibitory control (a central component of self-control) training on health behavior (Allom, Mullan, & Hagger, 2016; Jones et al., 2016). However, studies included in these meta-analyses typically measured the outcome variable(s) directly after the training, leaving the possibility of short term carry-over and demand effects on the outcome measurement. In addition, many studies employed training-specific outcomes (e.g., effects of training the inhibition of food-related reactions on subsequent eating behavior) while the current analysis focuses on far-transfer effects (i.e., practicing self-control in one domain and measuring effects in a different domain). In the studies included in the present analysis, these far-transfer effects were measured at least one day after the last training session. Thus, the overlap between these analyses and the present work is small due to the different aims and scopes.

2012; Fanelli, 2012), meta-analyses tend to overestimate population effect sizes (Ioannidis, 2008; Levine, Asada, & Carpenter, 2009).

Methods

The present review followed reporting guidelines for meta-analyses outlined in the PRISMA statement (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). The study was preregistered under the international prospective register of systematic reviews (PROSPERO; registration number: CRD42016033917, http://www.crd.york.ac.uk/prospero/). Following recent recommendations for the reproducibility of meta-analyses (Lakens, Hilgard, & Staaks, 2016) and to facilitate future updates of this work, we made all data, code, full documentation of our procedures, and additional supplementary analyses available on the Open Science Framework (https://osf.io/v7gxf/).

Inclusion criteria

Studies were eligible for inclusion if they (1) implemented at least one training procedure that contained the repeated control of dominant responses, (2) included at least one control group, (3) allocated participants randomly to conditions, (4) measured at least one self-control related outcome variable in a different domain than the domain the training occurred in, (5) assessed the outcome variable(s) at least one day after the last training session[3], and (6) included samples of mentally healthy adults. We decided to only include studies with random allocation to conditions because only random allocation allows for a causal interpretation of training effects. For studies that contained conditions and/or

[3] This criterion was added to exclude studies that measured dependent variables only directly after the last training session, raising the possibility of short term priming or demand effects. We made one exception from the rule for the following reasons: Lin and colleagues (Lin, Miles, Inzlicht, & Francis, 2016) measured various dependent variables *repeatedly* during a 30-day training period, but not after the training period. We decided to include this study for two reasons: First, the study did not employ specific training sessions that would open the window for short term priming and demand effects, but employed a training procedure that instructed participants to use their non-dominant hand for everyday life activities five days a week from 8am to 6pm. Second, the measurements a) took place in a non-formalized context (online at home) and several dependent variables did not assess behavior or experience specific to the moment of assessment; instead these outcome variables pertained to longer time spans (e.g., the previous week).

outcomes irrelevant to our research question, we only included the conditions and/or outcome variables that matched all criteria. In case of ambiguity about the relevance of the chosen outcome variable(s), we generally followed the arguments of the original study authors. For a detailed documentation of all decisions that were made, see the documentation available on the OSF.

Search strategy

We conducted a systematic literature search using three online citation-database providers, namely EBSCO, ProQuest and ISI Web of Science. In EBSCO we searched the data-bases PsycINFO, ERIC, PsycARTICLES and PSYNDEX, using the exact search term (TI self regulat* OR TI self control OR TI inhibit* OR TI willpower) AND (TI training OR TI practic* OR TI exercis* OR TI improv*). For ISI Web of Science, the exact search term was TITLE: ([self regulat* OR self control OR inhibition OR willpower] AND [training OR practic* OR exercis* OR improv*]). This search was restricted to entries tagged as "psychology". In ProQuest we searched for (["self regulat*" OR "self control" OR "inhibition" OR "willpower"] AND ["training" OR "practice" OR "exercise"] AND "psychology"). All data-bases were searched from 1999 onward, the publication year of the first self-control training study (Muraven, Baumeister, & Tice, 1999). Additionally, we issued calls for unpublished data through the mailing lists of three scientific societies (SPSP, EASP, SASP) and personally corresponded with researchers that are active in the field. Finally, the literature search was complemented by unsystematic searches and reference harvesting from included studies and relevant overview articles.

Screening

Titles and abstracts of 4,075 records were screened by the second author for relevance to the present work. Of these, 4,026 were excluded. Forty-nine full-text articles were assessed

for eligibility according to the inclusion criteria. Twenty-eight were included in the final database. The PRISMA flow chart in Figure 1 provides details about these steps.

Study coding

We coded several potential moderator variables of self-control training effects. One potential moderator pertains to the type of training that was implemented, some pertain to the study level, and some pertain to level of the outcome used. For further potential moderators and respective analyses, please see the supplement. The second author and a research assistant coded all potential moderators explained in the remainder of this section (see documentation on the OSF for details). Inter-rater-reliability was examined using intra-class correlation for continuous moderators (ICC[1,1]; Shrout & Fleiss, 1979) and Cohen's Kappa for categorical moderators (Cohen, 1968). Interrater reliability for the study coding was high by common standards (Cicchetti, 1994), mean $\kappa = 0.83$, mean $ICC(1,1) = 0.92$.

Treatment-level moderator

Type of training. Some training procedures may be more effective than others. For example, training procedures that require more deliberate and effortful behavioral control (e.g., repeatedly squeezing a handgrip over several weeks) may differ in effectiveness from training procedures that require more frequent but less rigorous behavioral control (e.g., using one's non-dominant hand for everyday activities).

Study-level moderators

Length of training. Longer training procedures may lead to stronger training effects. Length of treatment was coded in days. Length of training was coded as a study-level (instead of a treatment-level) moderator because in all studies with more than one treatment condition treatment length was equal across conditions.

Publication status. Studies with statistically significant results are more likely to be published, possibly leading to an overestimation of the average effect size. Published and *in*

press studies were coded as published, all others as unpublished. (For a more comprehensive treatment of potential publication bias, see below.)

Research group. The self-control training hypothesis was derived from the strength model of self-control (Baumeister et al., 2007). Perhaps researchers from this group are more experienced and more skilled at operationalizing relevant variables than other researchers. Alternatively, they may also be more biased in favor of the self-control training hypothesis. Given the criticisms to the strength model it is also possible that researchers from other research groups are biased against the hypothesis. Following Hagger et al. (2010), a study was coded 'Strength model research group' if one of the authors or committee members of a dissertation or Master's thesis was Roy Baumeister or one of his known collaborators (alphabetically: DeWall, Gailliot, Muraven, Schmeichel, Vohs). All other studies were coded 'other'.

Control group quality. Intervention effects that are based on comparisons of training conditions with inactive control groups can result from multiple different working mechanisms (e.g., demand effects, stronger engagement in the study in the intervention group, etc.). Active control groups narrow down the range of plausible working mechanisms and provide a more conservative test of the self-control training hypothesis. Control groups were coded as active when they worked on any task while the intervention group received treatment; all other control groups were coded as inactive.

Gender ratio. Meta-analytic evidence suggests that trait self-control is more strongly linked to the inhibition of undesired behaviors in males than in females (de Ridder, Lensvelt-Mulders, Finkenauer, Stok, & Baumeister, 2012). Thus, to the extent that self-control training improves trait self-control, training may show stronger effects in males than in females. We coded the gender ratio as the percentage of males in the sample.

Outcome-level moderators

Type of outcome. Training effects on some outcome variables may be stronger than on others. We grouped outcome variables into clusters representing different content domains (e.g., physical persistence, health behaviors, academic behaviors).

Lab versus real-world behavior. For some outcomes, the relevant behavior is performed in the laboratory (e.g., computerized performance tasks). For others, the relevant behavior refers to real-world behavior performed outside the laboratory (e.g., "How often have you done X during the last week?") and may also be assessed outside the laboratory (e.g., daily diaries). Behavior assessed in the laboratory may provide more experimental control, variables that reflect real-world behavior or experience may have higher external validity. Outcomes were coded as 'lab behavior' or 'real-world behavior'.

Stamina versus strength. Some outcomes were assessed without a preceding effortful task, others after an effortful task. Outcomes were coded as 'self-control stamina' (i.e., resistance to ego depletion) when they were preceded by an effortful task and as 'self-control strength' when they were not preceded by an effortful task.

Maximum versus realized potential. Some dependent variables require the participant to perform as well as possible (i.e., realize their full self-control potential; e.g., Stroop task or keep hand in ice water for as long as possible). When not prompted, people may not always access their maximum potential but realize only a part of it in a given situation. Self-control training may differentially affect the maximum potential people *can* exert and the realized potential they *do* willingly exert.

Follow-up. Training effects may deteriorate with increasing time between the end of training and outcome measurement. Follow-up was coded as the number of days between the last day of training and the outcome measurement. If the outcome measurement spanned across a period of time, the middle of this time period was used to calculate follow-up.

Effect size coding

We computed Hedges' *g* effect sizes and respective variances (Var_g) for all effects (Hedges, 1981). Hedges' *g* is similar to Cohen's *d*, but corrects for small sample bias. Two design types were prevalent: Pretest-posttest-control designs (PPC) and posttest-only-control designs (POC). For continuous dependent variables, we first computed Cohen's *d* and its variance Var_d and then applied Hedges' correction factor for small sample bias to compute *g* and Var_g. For PPC designs, Cohen's *d* was defined as the difference of mean improvement between the training group and the control group, divided by the pooled pretest standard deviation (*SD*):

$$d_{PPC} = \frac{(M_{Treat[POST]} - M_{Treat[PRE]}) - (M_{Ctrl[POST]} - M_{Ctrl[PRE]})}{\sqrt{\left((n_{Treat}-1)*SD^2_{Treat[PRE]} + (n_{Ctrl}-1)*SD^2_{Ctrl[PRE]}\right)*\frac{1}{n_{Treat}+n_{Ctrl}-2}}} \quad (1)$$

Thus, the numerator in the Cohen's *d* fraction was a difference of differences, that is, the difference of the mean improvement ($M_{post} - M_{pre}$) between the two conditions. Standardizing by pooled pretest *SD* rather than pooled posttest *SD* or pooled total *SD* has been shown to yield a more precise estimate of the true effect, as interventions typically cause greater variation at posttest (Morris, 2008).

For POC designs, Cohen's *d* was defined as the difference in means divided by the pooled posttest standard deviation.

$$d_{POC} = \frac{M_{Treat[POST]} - M_{Ctrl[POST]}}{\sqrt{\left((n_{Treat}-1)*SD^2_{Treat[POST]} + (n_{Ctrl}-1)*SD^2_{Ctrl[POST]}\right)*\frac{1}{n_{Treat}+n_{Ctrl}-2}}} \quad (2)$$

For non-continuous variables, appropriate effect sizes for the respective scale level were computed and then transformed to Hedges' *g* (Hedges, 1981). When possible, effect sizes were computed from descriptive statistics and sample sizes. We contacted the authors if required information was missing in the manuscript. Eighteen out of 23 responded to our inquiry. If authors did not respond or could not provide the required information, we approximated the effect size as closely as possible using the information provided in the original manuscript.

Some studies included more than one treatment group or control group (e.g., using self-control training tasks and/or control tasks from different domains). When multiple treatment and/or control groups were implemented, we compared each treatment group separately against each control group. For studies that included multiple outcomes, we computed one effect size per outcome for each comparison. For example, a study reporting two treatment groups, two control groups, and three outcomes would contribute a total of twelve effect sizes (2 treatments × 2 controls × 3 outcomes). Some studies reported multiple measurements of the same outcomes after training. In these cases, we only included the measurement temporally most proximate to the training phase (exception: follow-up moderator analysis, see next paragraph).

For the moderator analysis 'Follow-up', we contrasted outcome variables measured directly after the training (post-training, see above) with later measurement occasions (follow-up). If a study included both post-training and follow-up measurements, we included effect sizes for both time points. When multiple training and/or control groups were implemented, we combined them, respectively, before computing the effect sizes, since type of training/control group was not of interest in this particular analysis.

Meta-analytic procedure

We deviated from the path of data analysis outlined in the pre-registration because we followed valuable reviewer suggestions made in the editorial process (i.e. reliance on the robust variance estimation approach, see below). All analyses were conducted using random effects models because self-control training interventions, control groups, and outcome variables varied considerably between studies. Hence, it was unreasonable to expect one true, "fixed" population effect.

Conventional meta-analytical techniques assume that effect sizes are statistically independent. Including multiple effect sizes stemming from multiple outcomes or

comparisons per study violates this assumption (Lipsey & Wilson, 2001). Several approaches have been proposed to address this issue and to arrive at a set of independent effect sizes (for an overview, see Marín-Martínez & Sánchez-Meca, 1999). One widely used approach averages and adjusts effect sizes based on the correlation of the combined effect sizes (Borenstein, Hedges, Higgins, & Rothstein, 2009). More specifically, the effect size variance estimate is more strongly reduced if the combined outcomes are weakly correlated compared to when they are highly correlated. This reflects the idea that uncorrelated outcomes contain broader informational value than highly correlated outcomes. One downside of this approach is that averaging effect sizes leads to a loss of information because analyses on the level of effect sizes are no longer possible. To illustrate, consider a study reporting treatment effects on reading and mathematics achievement. Averaging these effect sizes delivers one study summary effect. The single summary effect prohibits a moderator analysis investigating effects of the treatment on different outcomes such as reading *versus* mathematic achievement across several studies in the meta-analysis.

The recently developed robust variance estimation (RVE) approach for meta-analysis (Hedges, Tipton, & Johnson, 2010) solves this issue. It permits conducting random effects meta-regression on dependent effect sizes, thus offering many advantages over the previously described averaging approach. Unfortunately, there are some drawbacks to RVE as well. First, RVE estimates the correlation matrix of dependent effect sizes, rather than accounting for it directly. It will therefore generally yield less precise results than approaches that incorporate the empirical correlation structure (e.g., the procedure proposed by Borenstein et al., 2009). Second, because the approach is relatively novel, the validity of some key meta-analytical techniques has not yet been validated in the RVE context, such as regression-based tests for small-study effects, trim-and-fill procedures, or power analyses. Third, while it is possible to calculate point estimates of true variance in the effect sizes in RVE (i.e., I^2), there

are currently no significance tests of these estimates available. Hence, researchers must rely on conventions when interpreting the true variance of effect sizes (Higgins & Green, 2011).

Considering the respective (dis)advantages of the Borenstein approach and the RVE approach, we adopted the following twofold strategy for the present meta-analysis: First, we computed the global summary effect of self-control training based on RVE and provide the parallel estimate based on the Borenstein approach for converging evidence. Second, all moderator analyses were run based on RVE. Third, all tests to detect and correct for small study effects were run based on the Borenstein approach, as the validity of these procedures has not yet been investigated in the RVE context. We also ran these analyses within the RVE approach for converging evidence. These latter analyses should be interpreted with caution, however. Please refer to the supplemental materials for details of the Borenstein approach. We relied on the *MAd* package to implement the approach (Del Re & Hoyt, 2014).

Robust Variance Estimation (RVE). All RVE models were fitted using the *robumeta* package for *R* (Fisher, Tipton, & Hou, 2016). We ran the RVE analyses with the following specifications: First, standard RVE has been shown to perform satisfactorily with a minimum of 10 studies when estimating summary effects, and with a minimum of 20-40 studies when estimating slopes in meta-regression (Hedges et al., 2010; Tipton, 2013). When the number of studies falls below these limits, significance tests tend to have inflated Type I error rates. We therefore implemented significance tests that incorporate small sample corrections for all RVE models (Tipton, 2015; Tipton & Pustejovsky, 2015). Specifically, we conducted Approximate Hotelling-Zhang tests for testing multiple parameters (Tipton & Pustejovsky, 2015, abbreviated HTZ in the *clubSandwhich* package that we employed to run these analyses, Pustejovsky, 2016), and *t*-tests for single parameters (Tipton, 2015). Both *HTZ* and *t*-values had small-sample-corrected degrees-of-freedom and adjusted variance-covariance matrices. It is important to note that the single-parameter *t*-test (but not the

multiple parameter *HTZ* test) may provide inaccurate results when degrees-of-freedom fall below $df = 4$. Consequently, we caution the reader to interpret the results when this was the case, and we refrained from reporting *p*-values and confidence intervals in the figures depicting analyses with dfs < 4.

Second, meta-analysts using RVE need to decide how to weight the effect sizes. Following recent recommendations (Tanner-Smith & Tipton, 2014), we set the weights to account for the type of dependence that is likely to be most prevalent in the dataset (i.e., dependence due to correlated rather than hierarchical effects). Third, we estimated the average correlation of effects sizes by first averaging all Fisher-*Z* transformed outcome correlations per study, averaging these means across all studies, and then transforming the value back to a Pearson correlation. This procedure returned a mean outcome correlation of $r = .18$. We additionally conducted sensitivity analyses for all models by varying the correlation estimate from $r = 0$ to $r = 1$ in steps of $r = .2$. This did not appreciably influence the conclusions drawn from the models. For example, the overall mean estimate of self-control training effectiveness only changed by $\Delta g = 0.0002$ when going from $r = 0$ to $r = 1$.

In order to compute the overall summary effect, we fitted an intercept-only random-effects RVE model to the set of dependent effect sizes. The regression coefficient of this model can be interpreted as the precision-weighted mean effect size of all studies, corrected for effect-size dependence. The corresponding significance test probes whether the estimate is significantly different from zero. To estimate the variance of true effects, we computed T^2 (DerSimonian & Laird, 2015), which estimates the true heterogeneity of effects in the same metric as the original effect size. For a more interpretable measure of heterogeneity, we also computed I^2 (Higgins, Thompson, Deeks, & Altman, 2003), which reflects the estimated proportion of true variance in the total observed variance of effect sizes.

To examine the convergence of RVE and the more conventional approach (Borenstein et al., 2009), we also computed an overall summary effect from the set of independent effect sizes by fitting a conventional random-effects model to the data using the *metafor* package (Viechtbauer, 2010, 2016). To test the dispersion of observed effect sizes for significance, we computed Cochran's *Q* (Cochran, 1954) that is defined as the ratio of observed variation to the within-study error. *Q* follows a χ^2 distribution. A significant *Q* value provides evidence that the true effects vary. We again computed T^2 and I^2 to estimate true heterogeneity. The summary effect was computed as the precision-weighted mean of all independent effect sizes. Weights were set to the inverse of the sum of the respective effect size variance (Var_g) and the estimated true heterogeneity (T^2).

Moderation analyses. To test for moderation, we employed mixed-effects RVE models. RVE offers the advantage that several moderators can be analyzed simultaneously while taking dependence of predictors (moderators) and outcomes (effect sizes) into account. These models logically extend standard multiple regression to meta-analysis. Accordingly, methodological concerns relevant in multiple regression are also relevant to meta-regression, especially overfitting of models, confounding among predictor variables, and low power. The number of studies and effect sizes was not large enough to include all coded moderators in a single model. We therefore followed a stepwise procedure to analyze the effect of moderators on the summary self-control training effect. In a first step, we separately tested the bivariate relationship of each moderator with the effect sizes. Categorical predictors were dummy coded, continuous predictors were entered without transformation. This step delivers evidence for moderators without accounting for the influence of other, potentially correlated moderators.

In a second step, we entered multiple predictors simultaneously into the model to control for possible confounds between moderators. To avoid overfitting of the model, it was

necessary to pre-select predictors. Since we had no a-priori theoretical rationale for the relative importance of the various moderators, we examined the converging evidence of a twofold strategy to determine the most suitable set of predictors. The first strategy was to select all moderators with *p*-values of .100 or smaller in the bivariate tests (see previous paragraph). The second strategy was to fit models for all possible combinations of predictor variables. From this set, we retrieved the 100 models that explained the largest amount of true heterogeneity in the effect sizes as indicated by I^2. Next, we scored the relative importance of each moderator according to the following rule: A moderator received a score of 100 if it was included in the best model (i.e., the model explaining the largest amount of true heterogeneity), a score of 99 if it was included in the second best model, and so forth. Scores per moderator were summed up to create indices of relative importance. Thus, the maximum importance score was 5,050 for a moderator that was included in all of the hundred most potent models. We then chose moderators to be included in the model based on their importance indices. This approach should be less susceptible to chance patterns in the data biasing the model than simply selecting the model with the single lowest I^2 because relative importance across multiple models is taken into account. We developed this method of selecting predictors based on the idea of all-subsets methods in multiple regression (Hocking, 1976), as there are currently no other methods for model building in meta-analysis available.

Small-study effects and publication bias

Publication bias results if studies with certain characteristics (e.g., significant effects, large effect sizes) are systematically more likely to be submitted for publication by authors and/or accepted for publication by journals than studies with non-significant or negligible effect sizes. If this happens, the published literature is not representative of the full body of research and overestimates the population effect size (Ioannidis, 2008). Publication bias is a

pervasive problem in the social sciences including psychology (Bakker et al., 2012; Franco, Malhotra, & Simonovits, 2014, 2016).

When a given literature is affected by publication bias, there will likely be a negative relationship between studies' effect sizes and their precision (or sample size): More precise studies with larger samples yield smaller effect sizes. This relationship is found in many meta-analyses (Levine et al., 2009). Small studies are more likely than larger studies to be excluded from the published literature due to non-significance or to be influenced by questionable data analysis methods that lead to significant findings at the cost of a factually increased Type I error (e.g., *p*-hacking; Simmons, Nelson, & Simonsohn, 2011). Therefore, several statistical methods to detect and correct for publication bias investigate the relation between effect size and precision. These assume that in an unbiased literature small studies (on average) should be no more likely to deliver strong effects than larger studies.

It is important to note that a negative relationship between effect size and precision may also result from unproblematic causes other than publication bias. For example, smaller studies may have used other populations that may be more strongly affected by the intervention. Further, it is possible that certain particularly effective interventions are more readily applied in small than in large studies. Also, experimental manipulations may be more rigorously (and therefore more effectively) applied in small than in large studies (Sterne et al., 2011). These kinds of small-study effects reflect true heterogeneity of effect sizes. This heterogeneity may be quantified and potentially explained by statistical analyses such as moderator analyses. Importantly, they are not a problematic sign of publication bias. In case of an empirically negative association of effect size and study precision, meta-analysts therefore need to reflect about possible reasons for this relationship with respect to the specific body of research that is being investigated.

We applied two methods to *detect* publication bias (Funnel plot, Egger's regression test) and two further methods to *correct for* publication bias (Trim and Fill; Precision Effect Estimation with Standard Error [PEESE]). In the way they have been developed and validated, these techniques require statistical independence, so we applied them to the set of independent effect sizes (Borenstein approach). However, the logic of Egger's regression test and PEESE can be readily extended to RVE. We report both approaches for these procedures, but caution is warranted in interpreting the RVE variants until the techniques have been thoroughly validated in RVE.

Funnel plot. A funnel plot provides a graphical depiction of the relation between effect size and study precision. Effect size is plotted on the x-axis and precision (as indicated by the standard error of the study effect size) on the y-axis with highest precision on top. Funnel plots feature a triangle that is centered on the empirical fixed effect estimate. The width of the triangle is 1.96 standard errors to either side such that 95% of studies would be expected to fall within the triangle in the absence of small study effects and heterogeneity. Studies are expected to spread symmetrically around the estimated effect and increasingly closer to the actual population effect as precision increases. Asymmetry of the funnel plot indicates small study effects that may be indicative of publication bias. Importantly, the funnel plot assumes homogeneous effect sizes, that is, all interventions share the same underlying population effect size. This is an assumption that is unlikely for research in the social sciences (Borenstein et al., 2009). Under the more realistic assumption of a random-effects model and true heterogeneity, funnel plots may overestimate small study effects and, ultimately, publication bias (Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006).

Egger's regression test. Egger's regression test investigates whether there is a statistically significant relationship between effect sizes and study precision. The currently advocated variant is a random-effects meta-regression of study effect size on study standard

error with an additive between-study error component (Sterne & Egger, 2005). A significant regression weight for the studies' standard error indicates the presence of small-study effects and potentially publication bias. Similar to other regression-based methods, Egger's regression test suffers from low statistical power when the number of studies is small (Kromrey & Rendina-Gobioff, 2006). The test also performs unsatisfactorily under conditions of heterogeneity. However, this downside is partly compensated for by the advantage that the approach can incorporate other study characteristics (that may account for heterogeneity). This allows investigating whether a possible relation between study precision (as indicated by the study standard error) and effect size remains significant after controlling for other potential influences on effect sizes (Sterne & Egger, 2005). Extending the idea of the test, we additionally investigated the relationship of effect size and standard errors in a mixed-effects RVE meta-regression with dependent effect sizes.

Trim and Fill. The *Trim and Fill* method (Duval & Tweedie, 2000a, 2000b) investigates asymmetry in a funnel plot. The algorithm removes extreme studies until the funnel plot is symmetric, yielding (in theory) an unbiased overall effect size estimate. It then imputes mirror images of the trimmed studies to estimate the correct variance of the overall distribution of studies. The *Trim and Fill* method suffers from the funnel plot's problematic assumption of truly homogeneous studies and a fixed effect size. In fact, simulation studies showed that *Trim and Fill* may even adjust for publication bias when factually none exists; reversely, it may adjust insufficiently when in fact publication bias is strong—especially when a few precise studies diverge from the overall meta-analytic estimate (Inzlicht et al., 2015; Moreno et al., 2009; Terrin, Schmid, Lau, & Olkin, 2003). Another problem is that the method assumes publication bias to be driven by weak effects, whereas indeed it is more likely that it is driven by statistical non-significance (Simonsohn et al., 2014). Large studies

with significant results but weak effects are more likely to be published than smaller studies with big, but non-significant, effects.

Precision Effect Estimation with Standard Error (PEESE). *PEESE* (Stanley & Doucouliagos, 2014) computes meta-regression in which the squared standard errors of the effect sizes (an indicator of precision) predict the effect sizes. If there is a significant relationship, this may indicate small study effects and potentially publication bias. The intercept of this regression line is thought to indicate the effect size of a "perfect" study with a standard error of zero that is used as an indicator of the bias-corrected overall meta-analytic effect size. Because *PEESE* is based on linear regression, it works best in meta-analyses with large numbers of studies. We fitted an additive error random-effects model to derive the intercept for *PEESE*.[4] Additionally, we extended the logic of this test to RVE and investigated the intercept in a mixed-effects RVE model that regressed (dependent) effect sizes on squared standard errors.

Results

Characteristics of included studies

The search identified 4,075 articles, of which 28 were eligible for inclusion, contributing a total of 33 studies. See Figure 1 for a PRISMA flow chart and the

[4] *PEESE* is often used together with a similar method called *Precision Effect Test* (*PET*; Stanley & Doucouliagos, 2014). Similar to Egger's regression test, *PET* uses the effect sizes' standard errors as predictors instead of the squared standard errors in case of *PEESE*. In Egger's regression test, the regression weight of the standard error predictor is interpreted. *PET* interprets the intercept as the bias-corrected true effect size. *PET* has been heavily criticized based on evidence that the algorithm performs particularly poorly and severely underestimates the true effect size under a range of conditions typical for social psychology (e.g., heterogeneity, small number of studies; Gervais, 2015, 2016; Inzlicht et al., 2015; Reed, 2015). We therefore refrained from using *PET* to correct for publication bias. Two other recently proposed methods to estimate true effect sizes in meta-analyses are *p-curve* and *p-uniform* (Simonsohn et al., 2014; van Assen, van Aert, & Wicherts, 2015). Both methods rely exclusively on significant and published effect sizes. Also, only one *p*-value per study may enter the computation. For the present meta-analysis, these rules would have led to a substantial loss of information, because a considerable part of effect sizes were non-significant and/or unpublished. In addition, many studies included more than one dependent variable, of which we could have included only one. Of the total of 158 effect sizes less than 20 would have been available for the computation of the effect size estimates based on *p-curve* and *p-uniform*. We therefore refrained from applying these methods.

documentation on the OSF for (a) a list of excluded studies with reasons for exclusion, and (b) full references for all included studies. Out of these 33 studies, 10 were unpublished as of December, 14, 2016. Publication dates ranged from 1999 to 2016 (*Mdn* = 2014). Self-control training was operationalized through a diverse set of training paradigms. For instance, participants were prompted to use their non-dominant hand for everyday tasks, to complete multiple sessions of computerized inhibitory control tasks, or to control their diet. The majority of training procedures lasted two weeks (*m* = 19 effect sizes). In total, the analysis included data from 2,616 participants who were on average 21.63 years old. The average total sample size per study was *n* = 79, comprising mostly student samples (*k* = 27) and females (M_{female} = 67%). A wide array of outcomes was used to measure self-control related constructs after (*k* = 16) or both before *and* after the training (*k* = 17). Nine studies also included a follow-up measurement. Training effects were predominantly evaluated through inhibitory control tasks (*m* = 18), or in the domains of physical persistence (*m* = 15), health behavior (*m* = 16), and affect and well-being (*m* = 29).

Main analyses

Outlier treatment. Initial examination of the data showed that no effect deviated markedly from the rest of the distribution (z_{min} = -2.80, z_{max} = 2.78). Leave-one-out analyses showed that sequential removal of each effect size, respectively, did not strongly influence the RVE point estimate or precision of the summary effect (Δg_{min} = -0.022, Δg_{max} = 0.021, ΔI^2_{min} = -2.09%, ΔI^2_{max} = 1.24%). We therefore did not replace any effect sizes for the RVE analyses.

Examination of the independent study-level effect sizes (Borenstein approach) showed that one study (Davisson, 2013; *g* = -0.67) deviated markedly from the rest of the distribution as indicated by several influence statistics (*z* = -2.61 [next closest: *z* = -1.27], $r_{student}$ = -3.53 [next closest: $r_{student}$ = -1.13], *DFFITS* = -0.45 [clearly detached from the

distribution], Cook's $D = 0.16$ [clearly detached from the distribution]). The study also had a strong influence on the heterogeneity estimate ($\Delta I^2 = -12.21\%$). We therefore replaced this outlier effect size with the next most extreme effect size ($g = -0.16$) for all analyses based on independent study-level effect sizes.

Summary effect. The RVE random-effects mean effect size of self-control training was $g = 0.30$, CI_{95} [0.17, 0.42], $p < .001$, a small to medium effect size according to the conventions by Cohen (1988). More than half of the variance in observed effects sizes was estimated to reflect true differences in effect sizes ($I^2 = 59.13\%$, $T^2 = 0.093$). According to common conventions, this amount of heterogeneity can be classified as moderate-to-substantial (Higgins et al., 2003).

We also computed a summary effect from the set of independent effect sizes by fitting a conventional intercept-only random-effects model (Borenstein approach). This analysis largely replicated the results of the RVE model in terms of the point estimate ($g = 0.28$, CI_{95} [0.19, 0.38], $p < .001$) and heterogeneity ($I^2 = 48.47\%$, $Q[32] = 62.10$, $p = .001$, $T^2 = 0.032$). Study statistics and results of this analysis are depicted in Figure S1.

Moderator Analyses

Descriptive statistics, confidence intervals and inferential statistics of all categorical moderator variables are provided in Table 1. Numbers of effect sizes per group (m) are provided in parentheses. Results of the meta-regressions for continuous moderators are provided in Table 2.

Treatment-level moderator

Type of training. Five types of training procedures were applied in at least five studies. The most effect sizes originated from studies that used repeated sessions of computerized inhibitory control training ($g = 0.21$, $m = 56$), followed by training procedures prompting participants to use their non-dominant hand for everyday tasks ($g = 0.42$, $m = 49$).

Other common procedures required participants to repeatedly press and squeeze a hand strength training device until failure (g = 0.37, m = 21), to continuously regulate their posture by sitting and walking upright (g = 0.23, m = 11), or to continuously regulate their diet (g = -0.01, m = 8). Despite substantial descriptive differences, the overall analysis between the sub-groups was not significant, $HTZ(7.37) = 1.11$, $p = .421$ (Figure 2).

Study-level moderators

Length of training. The majority of studies used a training procedure with a duration of two weeks (m = 19; 58%). Thus, there was little variability in training duration, precluding a meaningful test of this moderator. Consequently, there was no significant moderation effect of the length of the training duration, $b_1 = 0.003$, $t(4.01) = 0.44$, $p = .682$ (Figure S2).

Publication status. On average, effect sizes were almost three times larger for published (g = 0.37, m = 131) than for unpublished studies (g = 0.13, m = 27). This difference was close to conventional levels of statistical significance, $t(16.47) = 1.76$, $p = .098$ (Figure S3).

Research group. Significantly larger effects were found by the "strength model research group" (g = 0.51, m = 22) compared to other research groups (g = 0.22, m = 136), $t(12.53) = 2.49$, $p = .028$ (Figure 3).

Control group quality. Descriptively smaller effects were evident in studies with active control groups (g = 0.23, m = 79) compared to studies with inactive control groups (g = 0.43, m = 79). The difference was close to statistical significance, $t(20.79) = 1.73$, $p = .099$ (Figure S4).

Gender ratio. We imputed two missing values for this moderator by fitting the linear model based on all but the respective two effect sizes, and then entering the two effect sizes in the regression equation, thus predicting the missing values from the effect sizes. The moderating effect of the percentage of males in the study samples was close to statistical

significance, $b_1 = 0.008$, $t(13.27) = 2.02$, $p = .064$, such that Hedges' g was predicted to increase by $\Delta g = 0.08$ per ten percent more males in the sample (Figure 4). Percentages ranged from 0 to 64% across studies, so any interpretation of this slope should be limited to this range.

Outcome-level moderators

Type of outcome. In total, the included studies featured 94 unique dependent variables. We grouped these variables into theoretically homogeneous clusters. Note that degrees-of-freedom for significance tests of subgroup summary effects are dependent on the number of studies and effect sizes within the respective cluster. Significance tests are only interpretable when $df > 4$ (Tipton & Pustejovsky, 2015). Additionally, small clusters in subgroup analyses can bias tests of other clusters and the full model because they tend to increase imbalance in categorical predictors. Thus, it was necessary to exclude small clusters from the analysis to arrive at a model for which all parameters are interpretable. To do so, we sequentially removed clusters with the lowest degrees-of-freedom, until all degrees-of-freedom for the single parameter tests were four or larger. This procedure retained five outcome clusters in the final model. These were affect and wellbeing ($g = 0.30$, $m = 29$), inhibitory control ($g = 0.17$, $m = 18$), physical persistence ($g = -0.06$, $m = 15$), health behavior ($g = 0.12$, $m = 16$), and inhibitory control after depletion ($g = 0.48$, $m = 9$). The difference between these outcome clusters was not significant, $HTZ(10.40) = 1.55$, $p = .259$ (Figure 5).

Lab-based versus real-world behavior. Effect sizes for outcomes that were measured in the lab ($g = 0.32$, $m = 79$) were not significantly different from outcomes that reflect real-world behavior ($g = 0.23$, $m = 79$), $t(16.32) = -0.88$, $p = .392$ (Figure S5).

Stamina versus strength. Effects for outcomes that were preceded by an effortful task (stamina; $g = 0.60$, $m = 29$) were remarkably larger than for outcomes that were not preceded by an effortful task (strength; $g = 0.21$, $m = 129$), $t(17.52) = -2.84$, $p = .011$ (Figure 6).

Maximum versus realized potential. Whether outcomes reflected maximum self-control potential ($g = 0.30$, $m = 54$) or realized self-control potential ($g = 0.30$, $m = 104$) had no effect on effect sizes, $t(27.75) < 0.01$, $p = .997$ (Figure S6).

Follow-up. The distribution of the time-lags between the last day of the training and the time of outcome measurement was discontinuous with very large variance and therefore inept for a regression analysis. We therefore ran a categorical moderation analysis comparing post-test shortly after training with follow-up measurements (see section on effect size coding). The follow-up measurements took place *Mdn* = 9.5 days after the last day of training (*M* = 42, *SD* = 65, *Min* = 3.5, *Max* = 184). Outcome measures that were assessed directly after the training yielded descriptively larger effect sizes ($g = 0.31$, $m = 74$) compared to outcomes measured at later time-points ($g = 0.18$, $m = 28$). This difference was not significant, $t(9.69) = 1.12$, $p = .291$ (Figure S7).

Multiple moderators. Testing multiple moderators simultaneously allows estimating the unique moderating role of each predictor whilst controlling for the overlap with other moderators. For this analysis, it was necessary to select a subset of moderators in order to avoid overfitting the model. Several moderators had to be excluded a priori from this process (e.g., due to missing values or restricted variance; please see the supplement for a full list of excluded moderators and reasons for exclusion).

As outlined in the methods section, we employed two approaches to select the most appropriate moderators for this combined analysis: One approach relied on the findings from the bivariate moderator analyses, the second approach was a data-driven bottom-up approach seeking to explain a high degree of heterogeneity with a small number of predictors.

Results of the bivariate analyses suggested entering four moderators with *p*-values of $p = .100$ or smaller in the respective bivariate analysis into the combined model: *Control group quality*, *stamina versus strength*, *research group*, and *gender ratio*. The data-driven bottom-up approach delivered converging evidence: We fitted multi-moderator models for all possible combinations of predictor variables, resulting in $2^9 = 512$ models, and retrieved the 100 models that explained the greatest amount of true heterogeneity (i.e., reduction in I^2). Figure S8 reports the relative importance of the nine examined moderators and can be interpreted akin to a Scree plot in factor analysis. There was a relatively large gap in importance between the fifth (*gender ratio*) and sixth (*subjectivity of outcome measurement*) most important moderators—suggesting to enter the first five moderators in the combined analysis. Four of these five moderators match those identified in the bivariate analyses. *Maximum versus realized potential* emerged as an additional important moderator despite being far from significance in the bivariate analysis ($p = .996$). This suggests that this moderator binds residual variation in the other predictors and thereby contributes to explaining heterogeneity (suppression effect; Conger, 1974). In summary, the approach based on the bivariate analyses and the data-driven bottom-up approach provided converging evidence for the relevance of four moderators, and the latter approach unveiled the contribution of one additional moderator potentially acting as a suppressor variable.

The full model including all five predictors was significant, $HTZ(13.46) = 3.32$, $p = .036$ (Table 3). The model explained $\Delta I^2 = 13.87\%$ more true effect size variance than the intercept-only model. The moderator *stamina versus strength* again emerged as significant ($p = .027$). For *Research group* there still was a trend towards significance ($p = .097$). The *p*-value for *control group quality* was almost unchanged compared to the bivariate analysis ($p = .097$). By contrast, *gender ratio* did not border on significance anymore ($p = .169$). The alleged suppressor variable, *maximum versus realized potential*, was also not significant ($p =$

.248). These findings suggest that three of the four moderators that were at least marginally significant in the bivariate tests tended to explain unique portions of effect size heterogeneity, even when controlling for the influence of the other most potent moderators.

Note that in this regression, shared variance between predictors contributes to the overall model fit, but is not assigned to any predictor specifically. Hence, to the extent that a predictor has a causal claim for parts of the non-assigned shared variance, even non-significant predictors may be important for the overall model. Non-significance of predictors should therefore not be over-interpreted as indicating that this predictor is unimportant in explaining heterogeneity.

Small-study effects and publication bias

Funnel Plot. Visual inspection of the funnel plot for the set of independent effect sizes (i.e., Borenstein approach, not RVE) revealed that the effect sizes were relatively symmetrically distributed around the summary effect (Figure S9). For perfect symmetry, a set of studies with small-to-negative effect sizes and low precision was missing (see Trim-and-Fill below). Six studies fell out of the interval in that 95% of studies would be expected for any given level of precision. This analysis suggests a moderate degree of small-study effects and potentially publication bias.

Egger's regression test. The slope for the meta-regression of independent effect sizes on standard errors was significant, $b_{se} = 1.51$, $SE = 0.61$, $z = 2.49$, $p = .013$, indicating a significant funnel plot asymmetry. We additionally entered covariates to examine whether standard errors had unique predictive value beyond other moderators (Sterne & Egger, 2005). We considered all moderators that were included in the multiple-predictor model reported above but could only enter *gender ratio* and *research group.* For the remaining moderators several studies realized more than one moderator value, precluding this moderator from the analysis (e.g., featuring both an active and an inactive control condition). The effect of

standard errors remained significant when controlling for *gender ratio* and *research group*, b_{se} = 1.29, SE = 0.62, z = 2.08, p = .038. Thus, Egger's regression test suggests a significant degree of small-study effects and potentially publication bias.

The RVE equivalent of Egger's regression test showed a similar, yet non-significant relationship between standard errors and effect sizes, b_{se} = 1.37, SE = 0.80, t(15.15) =1.70, p = .109. After reducing heterogeneity by controlling for all five moderators from the multiple moderator analysis reported above, the effect of standard errors was clearly not significant anymore, b_{se} = 0.36, SE = 0.70, t(11.86) = 0.52, p = .614. Follow-up analyses revealed that the notable change to the standard-error-only model in the p-value was primarily due to the fact that effect sizes for self-control stamina (vs. strength) and effect sizes for inactive (vs. active) control groups tended to have greater standard errors. When these two moderators were not controlled for, the p-value of the standard error predictor remained largely unchanged compared to the standard-error-only model (p = .136).

Trim and Fill. After the previously reported *bias-detection* techniques, we turned to *bias-correction* techniques. The *Trim and Fill* method indicated that four studies were missing on the left of the mean meta-analytic effect size in order to obtain a fully symmetrical funnel plot (Figure 7). Imputing these studies and adding them to the model delivered a bias-corrected random-effects summary estimate of g = 0.24, SE_g = 0.051, CI_{95} = [0.14, 0.34], p < .001, that can be most adequately compared to the corresponding uncorrected summary effect size estimate based on independent effect sizes (g = 0.28). This analysis suggests a moderate degree of small-study effects and potentially publication bias.

PEESE. The meta-regression of independent effect sizes on squared standard errors was significant, b_1 = 3.41, p = .008. The intercept that is thought to reflect the unbiased true meta-analytic summary effect was close to statistical significance, b_0 = 0.13, SE_b = 0.07, CI_{95} = [-0.01, 0.27], z = 1.86, p = .064. This corrected estimate is less than half of the size of the

uncorrected summary effect (g = 0.30 based on RVE, g = .28 based on the Borenstein approach). The *PEESE* analysis suggests substantial small-study effects and potentially publication bias. Regressing dependent effect sizes on squared standard errors in an RVE mixed-effects model yielded a non-significant intercept, b_0 = 0.12, SE_b = 0.11, CI_{95} = [-0.12, 0.36], $t(16.31) = 1.08$, p = .295.

Summary. Both the funnel plot as well as Egger's regression test suggest that there are small-study effects in the dataset that may be indicative of publication bias. The *Trim and Fill* method delivered a moderately adjusted bias-corrected effect size estimate. By contrast, the bias-corrected *PEESE* estimate was less than half of the initial summary effect and only marginally significant. Extending the logic of Egger's regression test and *PEESE* to the RVE framework provided largely converging evidence, but the *PEESE* estimate for the summary effect was clearly non-significant. Taken together, all available evidence suggests that there are small-study effects that may at least partly reflect publication bias. Unfortunately, the severity of this bias is difficult to estimate based on currently available methods, especially because the available methods do not closely converge.

Discussion

The present meta-analysis summarized studies testing the hypothesis that practicing self-control in one domain will lead to benefits in self-control performance in other domains. A random-effects meta-analysis based on 33 studies, 158 effect sizes and more than 2,600 participants revealed an overall effect size of g = 0.30, CI_{95} [0.17, 0.42]. Three comparisons help putting this effect size into perspective: First, it ranges between a small (0.2) and a medium (0.5) effect size according to the conventions by Cohen (1988), gravitating more toward a small than to a medium effect. Second, the effect size found here is a little larger than half of the average effect size found in a meta-analysis of 302 meta-analyses of a broad range of psychological, educational, and behavioral treatments (d = 0.50, *Mdn* = 0.47; Lipsey

& Wilson, 1993). Third, the current effect size ranges between the fourth and fifth decile of effect sizes in social psychology according to a meta-analysis of 322 meta-analyses in social psychology that revealed a mean effect of d = .43 (*Mdn* = .37; Richard, Bond, & Stokes-Zoota, 2003). In sum, the present meta-analysis suggests that repeated practice improves self-control with an effect size that is somewhat smaller than common treatment effects in general and effects in social psychology in particular.

The analysis also revealed a moderate to high degree of heterogeneity with about 60% of the variance estimated to be due to real differences in effect sizes. What are the underlying moderators that account for these differences? Training effects were stronger when they were assessed after performing an initial demanding self-control task, thus reflecting self-control stamina, as compared to assessments without such an initial task (reflecting self-control strength). This finding suggests that self-control training effects may be more pronounced when self-control demands accumulate (i.e., ego depletion).

Effects were also stronger when proponents of the strength model were involved compared to those conducted exclusively by other researchers. The origin of this effect is unclear. Possibly, proponents of the strength model operationalized treatments and instructed participants in particularly effective ways. Alternatively, strength model proponents may have been biased in favor of the hypothesis, or other researchers may have been biased against the hypothesis.

Effects also tended to be stronger in studies with inactive control conditions. This finding is plausible considering that inactive control conditions allow all kinds of mechanisms to drive training effects while active control conditions narrow down the range of possible driving mechanisms. Finally, self-control training tended to be more effective in males than in females. One reason for this effect could be that men have stronger potentially problematic behavioral impulses, as has been suggested by previous research (Baumeister,

Catanese, & Vohs, 2001; de Ridder et al., 2012). Men may therefore profit more from improved self-control through self-control training.

In an analysis that examined the most potent moderators simultaneously, *stamina versus strength*, *control group quality*, and *research group* remained at least marginally significant moderators. Gender ratio was no longer significant. Finally, it is noteworthy that even the comprehensive multi-moderator model explained only a moderate amount of heterogeneity ($\Delta I^2 = 13.89\%$, remaining $I^2 = 45.24\%$). This suggests that we either missed plausible moderating factors or that the bulk of variance in effect sizes is study-specific and not systematic.

In the course of working on this meta-analysis, we learned about another team of researchers working on a non-peer-reviewed analysis focusing on the effectiveness of self-control training to change health behavior (Beames, Schofield, & Denson, in press). Their work is related to the present analysis since the databases overlap. Yet, there are notable differences between the two projects: they rely on different meta-analytic approaches (robust variance estimation vs. conventional random-effects meta-analysis), the calculation of effect sizes differs for some study designs, and they investigate different moderator variables. Despite these differences, it is noteworthy and re-assuring that both analyses arrive at similar estimates for the uncorrected mean effectiveness of self-control training ($g = .30$ in the present analysis vs. $g = 0.36$ in the work by Beames and colleagues).

Small-study effects and publication bias

The *Trim-and-Fill* method indicated a moderate degree of bias and delivered a corrected effect sizes estimate of $g = 0.24$. By contrast, *PEESE* indicated a much greater degree of bias and delivered an estimate of $g = 0.13$ that was not significantly different from zero. Note that an association between effect size and study precision (as detected by *Trim-and-Fill* and *PEESE*) can result from publication bias, *p*-hacking, and other biases; but, it

may also partly or completely be due to mundane reasons that cause small-study effects. For example, in the medical sciences, samples that are particularly receptive to an intervention due to a certain health condition may show particularly strong effect sizes. Such samples may also be difficult to recruit and therefore form smaller sample sizes than samples consisting of more readily available (and less susceptible) participants. Concerning the present data base, we were unable to come up with analogous mundane reasons for small-study effects in the self-control training literature. Given how the field worked for many years (e.g., difficulty to publish non-significant findings), we deem it likely that there is publication bias in the investigated literature, but the severity of this bias is difficult to estimate. This is because none of the currently available techniques performs consistently well under conditions typical for (social) psychological literatures including heterogeneity and publication bias (Gervais, 2015; Inzlicht et al., 2015). Thus, the degree to which the bias-corrected estimates are biased themselves is unknown.

Mechanisms underlying training effects

The present meta-analysis suggests that self-control training may lead to slight improvements in self-control in other domains. The strength model postulates that the repeated control of dominant responses strengthens the "self-control muscle" (Baumeister & Vohs, 2016b). This metaphor is vivid and descriptive, but it is of limited explanatory value for the observed effects because it does not specify the psychological mechanisms explaining training success. What do we know about mechanisms underlying training effects? One may approach this question from two perspectives. First, one may try to identify the crucial elements in a self-control training that make it effective. Second, one may think about the psychological processes that mediate self-control training effects.

The strength model claims that the repeated exertion of self-control by overcoming a dominant response is the driving "ingredient" of the self-control training. However, effect

sizes stemming from studies with inactive control groups were almost 50% larger than those from studies with active control conditions. In studies with inactive control groups, various mechanisms besides the repeated control of dominant responses can cause an intervention effect (e.g., demand effects, greater engagement with the study by the active intervention group). What is more, even in the subset of studies with active control groups, few control groups closely matched the training condition, allowing for other than the focal mechanism to drive training effects. Thus, the net training effect due to the control of dominant responses may still be smaller than indicated by the training effect obtained for the studies employing active control groups (g_{active} = 0.23, CI_{95} = [0.08, 0.39]).

With regard to the mediating psychological processes surprisingly little is known. Some studies investigated changes in self-efficacy, awareness of the concept of self-control, and implicit theories about willpower as possible mechanisms but did not find evidence for mediation (Job, Friese, & Bernecker, 2015; Klinger, 2013; Muraven, 2010a, 2010b). In one study, self-control training reduced academic effort avoidance in university students, which partly mediated the effect of training on participants' grade point average (Job et al., 2015). This study suggests that motivational variables might play a mediating role. Future research has to test whether changes in effort avoidance may account for training effects in other domains than academic achievement.

One hitherto unexplored possibility is that training and control conditions differentially affect participants' expectations, thus allowing for placebo effects without actual changes in the trained constructs (Boot, Simons, Stothart, & Stutts, 2013; Foroughi, Monfort, Paczynski, McKnight, & Greenwood, 2016). Expectations regarding possible improvements on the dependent variables may differ between groups if they are not measured or, better, experimentally controlled – even in studies with active control groups. Hence, more knowledge is needed about how participants believe the (training or control)

intervention is affecting them. What do participants believe their training regimen to be good for? What are their ideas about the researchers' goals for the study, and which expectations about improvement on the measured constructs do participants hold?

In sum, little is known about the crucial elements of a training intervention. The literature to date does not deliver conclusive evidence that exerting self-control by repeatedly overriding dominant responses is the dominant *causal* mechanism that improves self-control over time and across domains. Even less is known about the psychological processes that are affected by a self-control training and lead to improved self-control performance.

How to move forward?

We will briefly discuss recommendations for future work concerning both methodological and theoretical developments. On the methodological level, future research should, first, conduct direct, high powered, and pre-registered replications. The set of the present 33 studies is very diverse, containing no close replications that would bolster confidence in obtained findings. Second, it will be important to more consistently use pre-post designs to increase statistical power. Based on the mean parameters evidenced by the current meta-analysis (g = .30, $N_{average}$ = 79, α = .05, $r_{pre\text{-}post}$ = .70 within control groups), power for studies with pre-post designs is adequate (1-β = .92). However, in post-only designs the same parameters result in a poor power of 37%, even with a one-tailed test. Note that it is possible that the true training effect is smaller than g = .30, which further increases demands on sample size. Third, future studies should employ (a) longer, and (b) more varying training durations as well as (c) more consistently include follow-up measurements with (d) varying time lags. Only nine of the analyzed 33 studies included a follow-up measurement (median time lag 9.5 days). Effect sizes post training were considerably larger (g = 0.31) than at follow-up (g = 0.18). Although non-significant, this difference raises concerns about the practical utility of self-control training in the way it has been implemented

to date. Researchers may want to consider ways to foster more sustainable self-control training, for example, by reminding participants of the training principles or implementing brief training refreshments after the main training period.

On the theoretical level, self-control training should only lead to performance improvements in activities that actually require self-control for a given person. This is not the case if a person has no goal to control a behavior. In this case, enacting such behaviors does not constitute a self-control failure. People who strive to achieve a certain goal or change a specific behavior – but are unsatisfied with their success in doing so (e.g., alcohol or nicotine consumption, eating behavior) – are the ones who are most likely to profit from a self-control training. For these people, a self-control training may constitute a welcome means to work on the goal and provide a motivational boost by conveying the possibility that the training may help to achieve the respective goal (even if the person has no elaborate idea about how the training may do so). Ideally, a training sets in motion recursive motivational processes that help to build and keep up adaptive routines that may then contribute to lasting changes in behavior (Walton, 2014).

In addition, it will be important to control for differences in expectations about the consequences of a training regimen because different expectations may factually drive training effects (Boot et al., 2013). Such placebo effects are interesting in their own regard, but they limit researchers' ability to draw causal conclusions about a proposed working mechanism of *self-control* training. However, from the perspective of people who are interested in self-control improvements, making progress toward goal attainment is more pressing than identifying the underlying processes. If placebo effects do the trick and do so reliably, one may pragmatically advocate to let them do it. Researchers may interpret such, at first, poorly understood effects as an opportunity to investigate the underlying (motivational) processes in depth and apply this knowledge to new training interventions.

Limitations

The present work suffers from some limitations that future research may want to ameliorate. First, with 33 studies the available evidence on self-control training is still moderate. In light of the analyses presented here, it is premature to draw far-reaching conclusions. Several moderator analyses delivered substantial descriptive differences that did not reach significance, potentially due to low power.

A second limitation is that we could not calculate publication bias-corrected effect size estimates to the extent we had initially planned. Some techniques proved very unsatisfactory in simulation studies in that they severely underestimated true effect sizes under almost all realistic conditions (PET; Gervais, 2015; Inzlicht et al., 2015). Several other recently introduced techniques appear promising (Simonsohn et al., 2014; van Assen et al., 2015), but cannot be applied in a reasonable way to the current literature. These procedures rely exclusively on significant and published effect sizes with only one reported *p*-value per study entering the computation. For the present meta-analysis, this would have led to an excessive loss of information (see footnote 5). Also, they assume a homogeneous distribution of effect sizes, an assumption clearly not valid in the present literature. Future developments in meta-analytic techniques may be able to deliver valid publication bias-corrected effect size estimates for literatures with similar characteristics as the present one.

Conclusion

Self-control is a domain-general capacity. The self-control training hypothesis suggests that practicing self-control in one domain improves self-control in other domains as well. The present random-effects meta-analysis found a small-to-medium self-control training effect. Bias-corrected estimates indicate a smaller effect. The working mechanisms underlying these far-transfer training effects are poorly understood and require further

attention. We hope this meta-analysis will inspire researchers to further engage in this theoretically intriguing and practically relevant field of psychological research.

References

Allom, V., Mullan, B., & Hagger, M. S. (2016). Does inhibitory control training improve health behaviour? A meta-analysis. *Health Psychology Review, 10*, 168-186.

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

Baumeister, R. F., Catanese, K. R., & Vohs, K. D. (2001). Is there a gender difference in strength of sex drive? Theoretical views, conceptual distinctions, and a review of relevent evidence. *Personality and Social Psychology Review, 5*, 242-273.

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. (in press). 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*. Abingdon, UK: Routledge.

Berkman, E. T. (2016). Self-regulation training. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (3rd ed., pp. 440-457). New York, NY: Guilford.

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8*, 445-454.

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

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.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101-129.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. 2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34*, 35-46.

Cranwell, J., Benford, S., Houghton, R. J., Golembewksi, M., Fischer, J. E., & Hagger, M. S. (2014). Increasing self-regulatory energy using an internet- based training application delivered by smartphone technology. *Cyberpsychology Behavior and Social Networking, 17*, 181-186.

Daly, M., Delaney, L., Egan, M., & Baumeister, R. F. (2015). Childhood self-control and unemployment throughout the life span: Evidence from two British cohort studies. *Psychological Science, 26*, 709-723.

Davisson, E. K. (2013). *Strengthening self-control by practicing inhibition and initiation.* Unpublished dissertation thesis, Duke University, Durham, NC. Retrieved from http://dukespace.lib.duke.edu/dspace/handle/10161/7258 on 11/06/2015

de Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review, 16*, 76-99.

Del Re, A. C., & Hoyt, W. T. (2014). MAd: Meta-Analysis with Mean Differences. R package version 0.8-2 [Computer software]. Retrieved from http://cran.r-project.org/web/packages/MAd

Denson, T. F., Capper, M. M., Oaten, M., Friese, M., & Schofield, T. P. (2011). Self-control training decreases aggression in response to provocation in aggressive individuals. *Journal of Research in Personality, 45*, 252-256.

DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials, 45*, 139-145.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16*, 939-944.

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

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

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904.

Finkel, E. J., DeWall, C. N., Slotter, E. B., Oaten, M., & Foshee, V. A. (2009). Self-regulatory failure and intimate partner violence perpetration. *Journal of Personality and Social Psychology, 97*, 483-499.

Fisher, Z., Tipton, E., & Hou, Z. (2016). robumeta: Robust Variance Meta-Regression. R package version 1.8 [Computer software]. Retrieved from https://cran.r-project.org/package=robumeta

Foroughi, C. K., Monfort, S. S., Paczynski, M., McKnight, P. E., & Greenwood, P. M. (2016). Placebo effects in cognitive training. *Proceedings of the national Academy of Sciences, 113*, 7470-7474.

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.

Gervais, W. M. (2015, June 16). Putting PET-PEESE to the test... Retrieved from http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1

Gervais, W. M. (2016, March 3). heavy PETting. Retrieved from http://willgervais.com/blog/2016/3/3/enough-heavy-petting

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

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.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312*, 1900-1902.

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

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65.

Higgins, J. P. T., & Green, S. (2011). Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated March 2011]. Retrieved from http://handbook.cochrane.org/

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557-560.

Hocking, R. R. (1976). Analysis and selection of variables in linear regression. *Biometrics, 32*, 1-49.

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.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640-648.

Job, V., Friese, M., & Bernecker, K. (2015). Effects of practicing self-control on academic performance. *Motivation Science, 1*, 219-232.

Jones, A., Di Lemma, L. C. G., Robinson, E., Christiansen, P., Nolan, S., Tudur-Smith, C., et al. (2016). Inhibitory control training for appetitive behaviour change: A meta-analytic

investigation of mechanisms of action and moderators of effectiveness. *Appetite, 97*, 16-28.

Klinger, J. (2013). *Examining mechanisms of self-control improvement*. Unpublished master's thesis. University of Waterloo, Waterloo.

Kromrey, J. D., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement, 66*, 357-373.

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

Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). Evidence based medicine: The case of the misleading funnel plot. *British Medical Journal, 333*, 597-600.

Levine, T., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs, 76*, 286-302.

Lin, H., Miles, E., Inzlicht, M., & Francis, Z. (2016). *Mechanisms underlying self-control training*. Manuscript in preparation.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect-sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2*, 32-38.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*, 270-291.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science, 11*, 512-534.

Miles, E., Sheeran, P., Baird, H., Macdonald, I., Webb, T. L., & Harris, P. R. (2016). Does self-control improve with practice? Evidence from a six-week training program. *Journal of Experimental Psychology: General, 145*, 1075-1091.

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., et al. (2011). 'Willpower' over the life span: decomposing self-regulation. *Social Cognitive and Affective Neuroscience, 6*, 252-256.

Mischel, W., & Baker, N. (1975). Cognitive appraisals and transformations in delay behavior. *Journal of Personality and Social Psychology, 31*, 254-261.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*, 8-14.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the national Academy of Sciences, 108*, 2693-2698.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Plos Medicine, 6*, 7.

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.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 364-386.

Muraven, M. (2010a). Building self-control strength: Practicing self-control leads to improved self-control performance. *Journal of Experimental Social Psychology, 46*, 465-468.

Muraven, M. (2010b). Practicing self-control lowers the risk of smoking lapse. *Psychology of Addictive Behaviors, 24*, 446-452.

Muraven, M., Baumeister, R. F., & Tice, D. M. (1999). Longitudinal improvement of self-regulation through practice: Building self-control strength through repeated exercise. *Journal of Social Psychology, 139*, 446-457.

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

Oaten, M., & Cheng, K. (2006b). Longitudinal gains in self-regulation from regular physical exercise. *British Journal of Health Psychology, 11*, 717-733.

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

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., et al. (2010). Putting brain training to the test. *Nature, 465*, 775-778.

Piquero, A. R., Jennings, W. G., Farrington, D. P., Diamond, B., & Gonzalez, J. M. R. (2016). A meta-analysis update on the effectiveness of early self-control improvement programs to improve self-control and reduce delinquency. *Journal of Experimental Criminology, 12*, 249-264.

Pustejovsky, J. (2016). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package version 0.2.1.9000 [Computer software]. Retrieved from https://github.com/jepusto/clubSandwich

Reed, W. R. (2015). A Monte Carlo analysis of alternative meta-analysis estimators in the presence of publication bias. *Economics, 9*, 1-40.

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

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*, 628-654.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

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.

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.

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. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99-110). New York, NY: Wiley.

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.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72*, 271-324.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5*, 13-30.

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.

Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods, 4*, 169-187.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*, 375-393.

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*, 604-634.

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.

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

Viechtbauer, W. (2016). metafor: Meta-analysis package for R. R package version 1.9-9. [Computer software]. Retrieved from https://cran.r-project.org/package=metafor

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science, 23*, 73-82.

Table 1

Results of moderation analyses for categorical moderators.

	Summary Effect and 95% Confidence Interval								Test of Moderation			
Moderator	*g*	LL	UL	*t*	*df*	*p*	k_{study}	$m_{effects}$	*Statistic*	*df*	*p*	I^2
Treatment Level Moderator												
Type of Training									*HTZ* = 1.11	7.37	.421	54.85%
Inhibitory Control Task	0.21	-0.02	0.44	2.04	9.41	.070	11	56				
Handgrip	0.37	-	-	5.21	3.66	-	5	21				
Non-Dominant Hand	0.42	0.25	0.59	5.58	9.25	< .001	11	56				
Posture Regulation	0.23	-	-	2.55	2.53	-	4	11				
Diet Regulation	-0.01	-	-	-0.02	2.61	-	4	28				
Study Level Moderators												
Publication Status									*t* = 1.76	16.47	.098	56.48%
Published	0.37	0.24	0.51	5.83	20.53	< .001	23	131				
Unpublished	0.13	-0.16	0.41	1.01	8.52	.338	10	27				
Research Group									*t* = 2.49	12.53	.028	55.61%

Strength Model	0.51	0.29	0.74	5.42	7.20	< .001	9	22				
Other	0.22	0.08	0.36	3.19	21.81	.004	24	136				
Control Group Quality									t = 1.73	20.79	.099	57.43%
Active Control Group	0.23	0.08	0.39	3.10	19.70	.006	22	79				
Inactive Control Group	0.43	0.23	0.64	4.68	11.02	< .001	13	79				
Outcome Level Moderators												
Type of Outcome									*HTZ* = 1.55	10.40	.259	62.76%
Affect and Well-Being	0.30	-0.12	0.71	1.87	4.70	.124	6	29				
Health Behavior	0.12	-0.21	0.45	1.01	4.01	.368	6	16				
Inhibition	0.17	-0.26	0.59	0.90	8.30	.395	11	18				
Inhibition after Ego Depletion	0.48	0.10	0.86	3.33	4.54	.024	6	9				
Physical Persistence	-0.06	-0.42	0.29	-0.46	5.28	.665	8	15				
Subjectivity of Outcome Measurement									t = 0.30	26.07	.588	59.79%
Other	0.32	0.13	0.51	3.50	21.90	.002	26	80				
Subjective	0.26	0.14	0.39	4.44	13.93	< .001	18	78				
Lab-based versus Real-World Behavior									t = -0.88	16.32	.392	59.35%

	Lab-Based	0.32	0.16	0.48	4.18	24.35	< .001	29	79			
	Real-World	0.23	0.05	0.40	2.93	10.00	.015	12	79			
Stamina versus Strength										t = -2.84	17.52	.011	56.50%
	Stamina	0.60	0.33	0.87	4.83	11.79	< .001	16	29			
	Strength	0.21	0.07	0.34	3.14	23.92	.004	28	129			
Maximum versus realized Potential										t < 0.01	27.75	.997	59.36%
	Maximum	0.30	0.02	0.58	2.26	15.91	.038	21	54			
	Realized	0.30	0.19	0.40	5.91	19.74	< .001	23	104			
Follow-up										t = 1.12	9.69	.291	61.22%
	Follow-Up	0.18	-0.02	0.39	2.16	6.74	.069	28	9			
	Post Training	0.31	0.16	0.45	4.32	27.00	< .001	74	31			

Note. g = effect size. LL = lower limit of the 95% confidence interval (CI). UL = upper limit of the 95% CI. t = t-value associated with the g-value in the same row testing statistical significance in the respective moderator level. *df* = associated small sample corrected degrees-of-freedom. p = p-value associated with the t-value and *df* in the same row. k_{study} = number of studies that that contributed to the respective moderator level. m_{effect} = number of effect sizes in the respective moderator category. *Statistic* (test of moderation): t-value for single parameter tests or Hotelling-T-approximated (*HTZ*) test statistic for multiple parameter tests. Significant test statistics indicate significance of the overall model. I^2 = reflects the proportion of true variance in the total observed variance of effect sizes after accounting for the respective moderator. For some moderator models the values for I^2 can become larger than for the global summary-effect model because of missing values or differences in effect size computation. Note that for three subgroups in the *type of training* analysis, degrees-of-freedom fell below 4. Significance tests for the summary effects should thus not be interpreted. Accordingly, we did not report CI_{95} and p-values for the respective subgroups.

Table 2
Results of moderation analyses for continuous moderators.

	Meta-Regression		Test of Moderation			
Moderator	Intercept	Slope	*t*	*df*	*p*	I^2
Study Level Moderators						
Length of Training	0.25	0.003	0.44	4.01	.682	60.58%
Gender Ratio	0.04	0.008	2.02	13.27	.064	55.83%

Note. Test of Moderation: *t*-value and corresponding small-sample corrected degrees-of-freedom. Significant *t*-values indicates significant moderation. I^2 = reflects the proportion of true variance in the total observed variance of effect sizes. For some moderator models the values for I^2 can become larger than for the global summary-effect model because of missing values or differences in effect size computation.

Table 3
Summary of RVE mixed-effects meta-regression model predicting effect sizes from multiple moderators.

Variable	*b*	*SE(b)*	*t*	*df*	*p*
Intercept	0.175	0.169	1.04	13.57	.317
Control Group Quality (Inactive)	0.207	0.116	1.78	16.25	.094
Stamina versus Strength (Stamina)	-0.387	0.155	-2.50	13.20	.027
Research Group (Strength Model)	0.205	0.114	1.80	12.17	.097
Self-Control Potential (Realized)	0.174	0.146	1.20	16.84	.248
Gender Ratio	0.006	0.004	1.45	13.71	.169

Note. Categorical predictors were dummy coded with 0 and 1. The moderator level coded as 1 is indicated in parentheses. *b* = regression coefficient. *SE(b)* = standard error of regression coefficient. *t* = *t*-value testing whether the regression coefficient in the same row is significantly different from zero. df = corresponding small-sample corrected degrees-of-freedom. *p* = *p*-value associated with the *t*-value and *df* in the same row. The full model was significant, $HTZ(13.46) = 3.32$, $p = .036$, $I^2 = 45.24\%$.

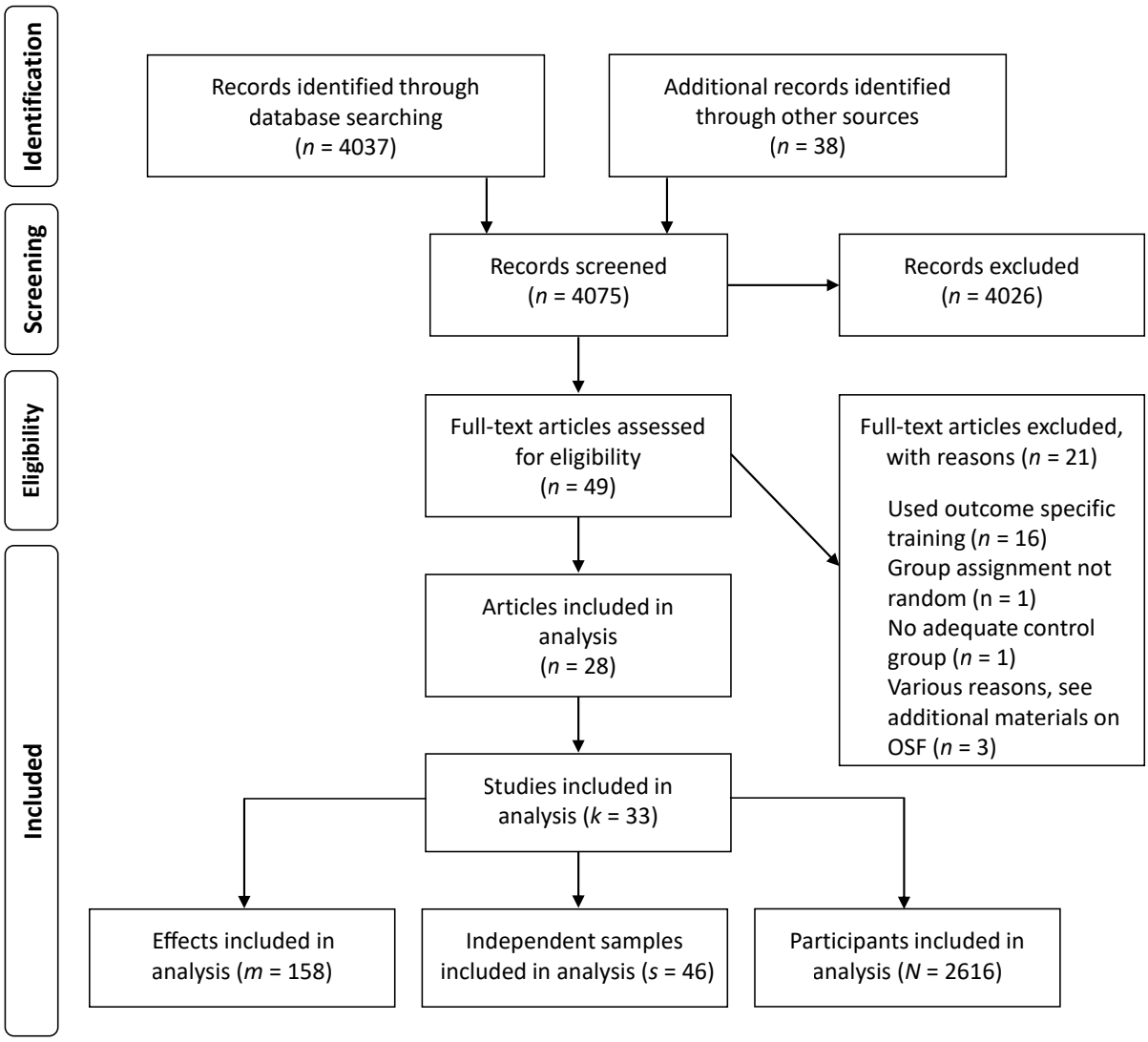

Figure 1. PRISMA flow chart of the literature search and study coding.

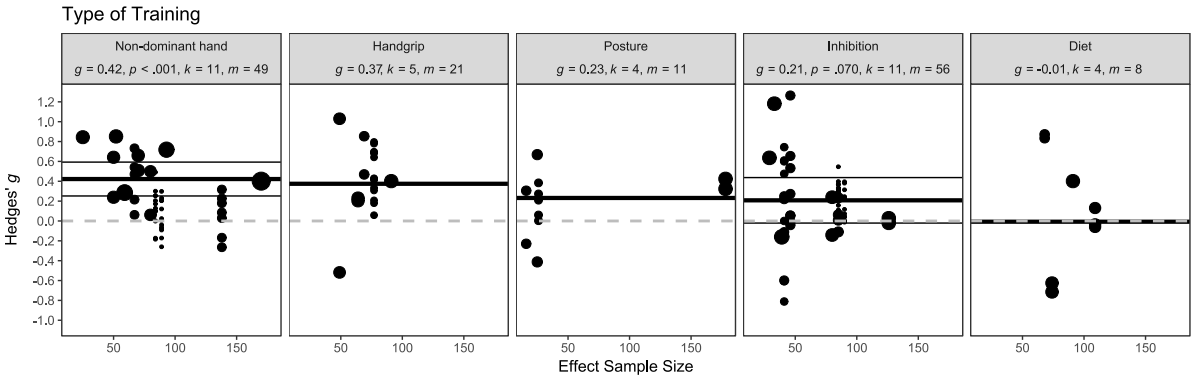

Figure 2. Moderation by type of training (*HTZ*[7.37] = 1.11, p = .421). g = Hedges' g summary effect within the respective subgroup. p = p-value testing Hedges' g against zero. k = number of studies in a subgroup. m = number of effect sizes in a subgroup. Black dots represent individual effect sizes. The thick black horizontal lines represent the meta-analytic summary effects within the subgroups. The thin black horizontal lines represent the borders of the 95% confidence interval around the subgroup summary effect. The dashed grey horizontal line represents the null-effect at g = 0. For informational purposes, the sample size that was used to calculate the respective effect size is depicted on the x-axis, but the moderating role of this attribute is not investigated in this analysis. Circle size represents the weight of the respective effect size in the meta-analytic RVE mixed-effects model depicted here. Non-dominant hand: Use of non-dominant hand for everyday tasks. Handgrip: Repeated use of a handgrip squeezer. Posture: Keep an upright posture in everyday life. Inhibition: Computerized inhibition control training procedures. Diet: Control one's diet. Note that for three subgroups in this analysis, degrees-of-freedom fell below 4. The corresponding significance tests for the summary effects should thus not be interpreted. Accordingly, we did not report CI_{95} and p-values for the respective subgroups.

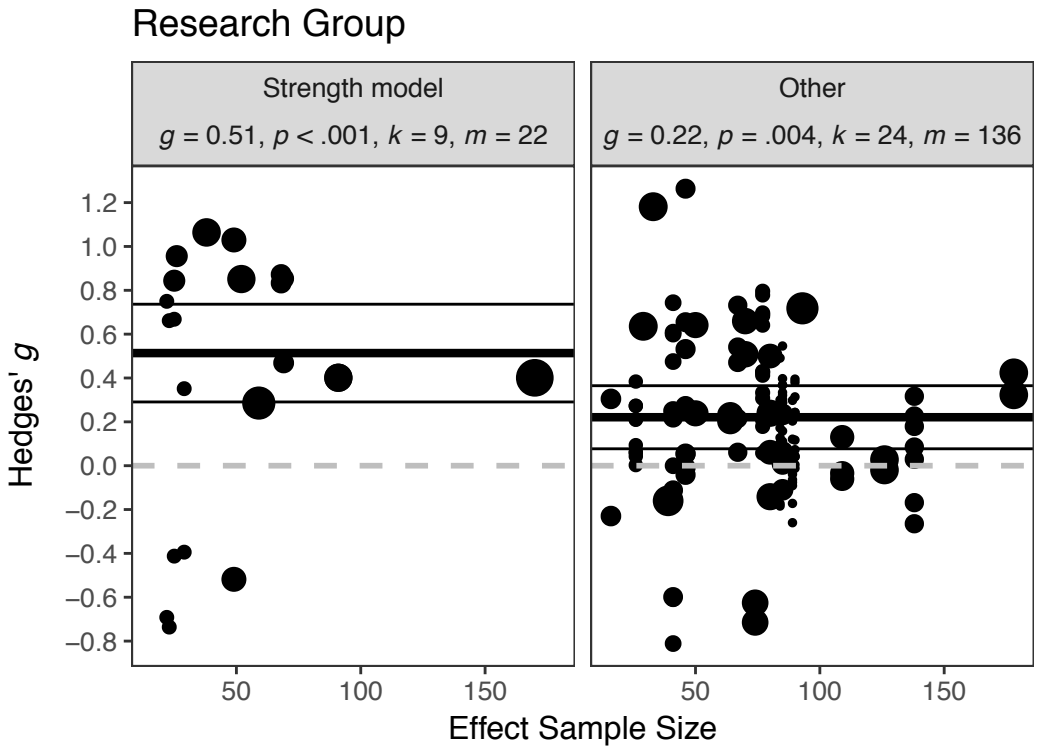

Figure 3. Moderation by research group (t[12.53] = 2.49, p = .028). g = Hedges' g summary effect in subgroup. p = p-value testing Hedges' g against zero. k = number of different studies within subgroup. m = number of effect sizes within subgroup. Black dots represent individual effect sizes. Thick black horizontal line: Meta-analytic summary effect within subgroup. Thin black lines: 95% confidence interval. Dashed grey line: Null-effect at g = 0. The associated sample size for each effect size is depicted on the x-axis for informational purposes. Circle size represents effect size weight for the subgroup analysis.

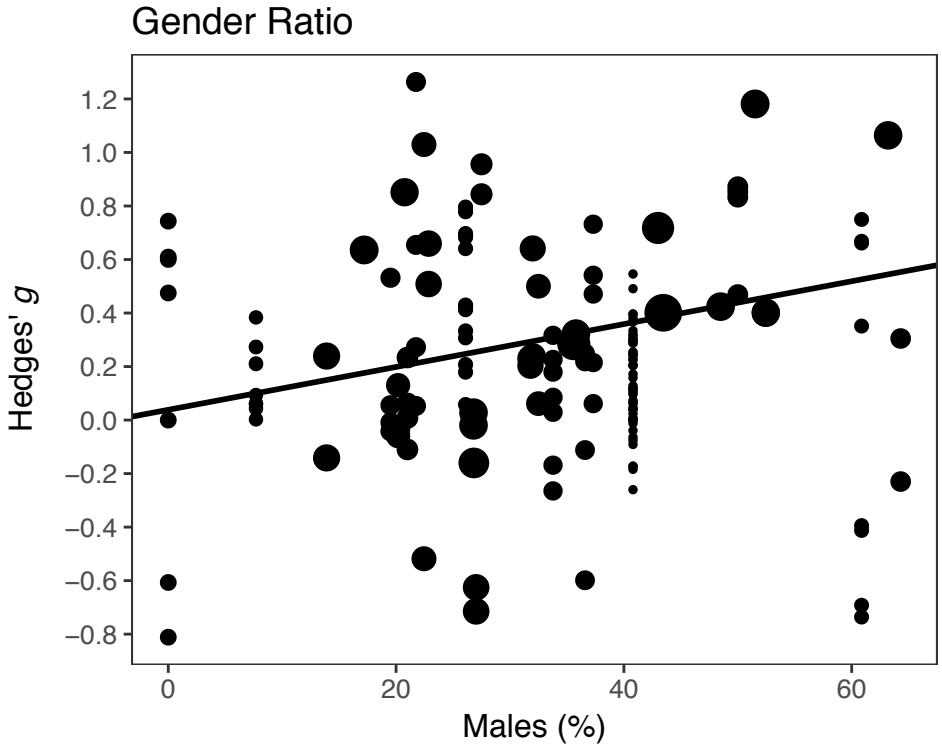

Figure 4. Moderation by gender ratio. The line represents the weighted RVE meta-regression of effect size on gender ratio (b_1 = 0.008, t[13.27] = 2.02, p = .064). Circle size represents effect size weight.

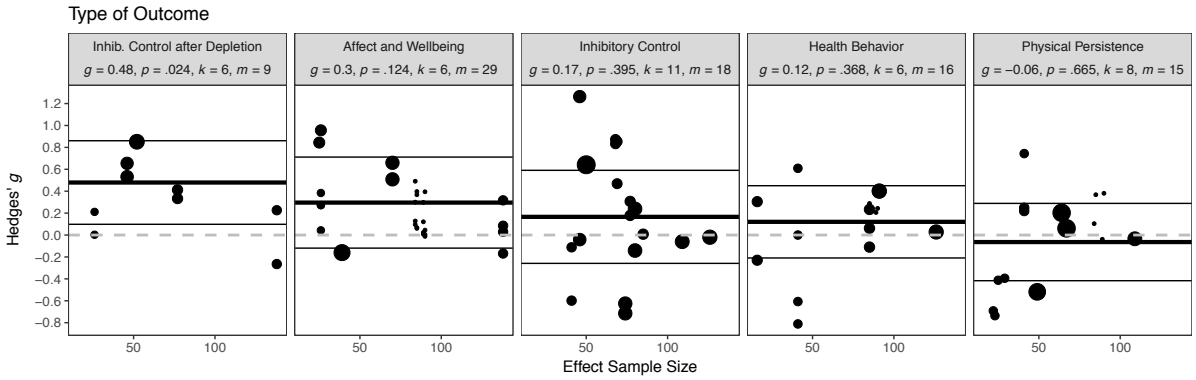

Figure 5. Moderation by type of outcome (*HTZ*[10.40] = 1.55, *p* = .259). *g* = Hedges' *g* summary effect in subgroup. *p* = *p*-value testing Hedges' *g* against zero. *k* = number of different studies within subgroup. *m* = number of effect sizes within subgroup. Black dots represent individual effect sizes. Thick black horizontal line: Meta-analytic summary effect within subgroup. Thin black lines: 95% confidence interval. Dashed grey line: Null-effect at *g* = 0. The associated sample size for each effect size is depicted on the *x*-axis for informational purposes. Circle size represents effect size weight for the subgroup analysis.

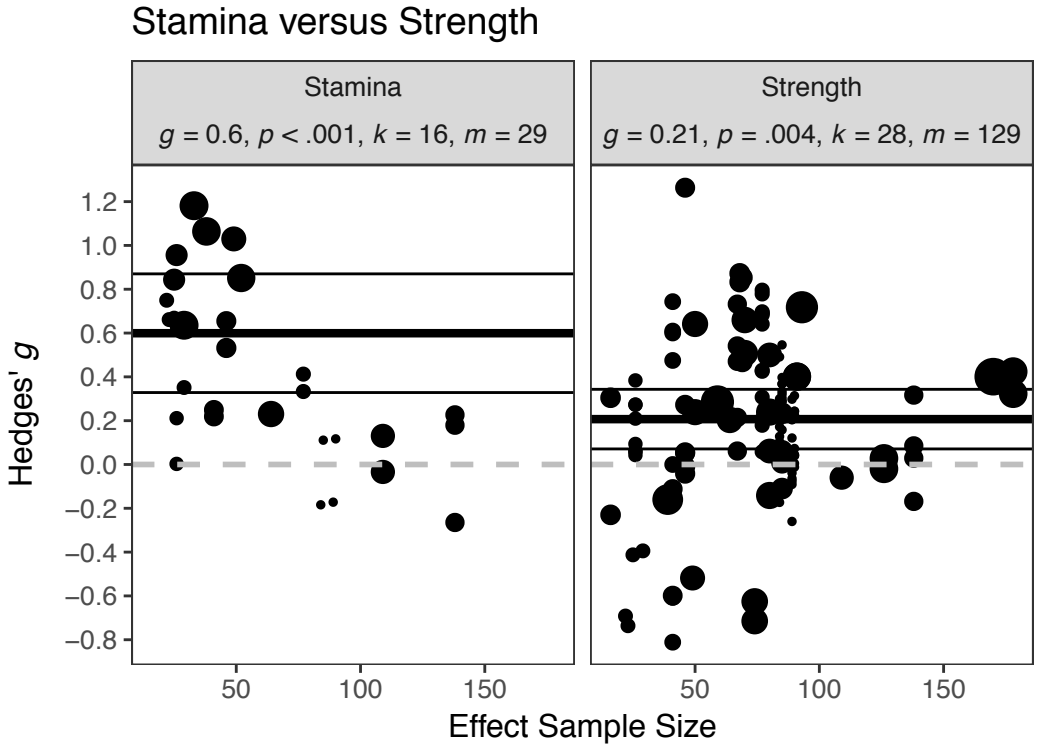

Figure 6. Moderation by strength versus stamina (t[17.52] = -2.84, p = .011). g = Hedges' g summary effect in subgroup. p = p-value testing Hedges' g against zero. k = number of different studies within subgroup. m = number of effect sizes within subgroup. Black dots represent individual effect sizes. Thick black horizontal line: Meta-analytic summary effect within subgroup. Thin black lines: 95% confidence interval. Dashed grey line: Null-effect at g = 0. The associated sample size for each effect size is depicted on the x-axis for informational purposes. Circle size represents effect size weight for the subgroup analysis.

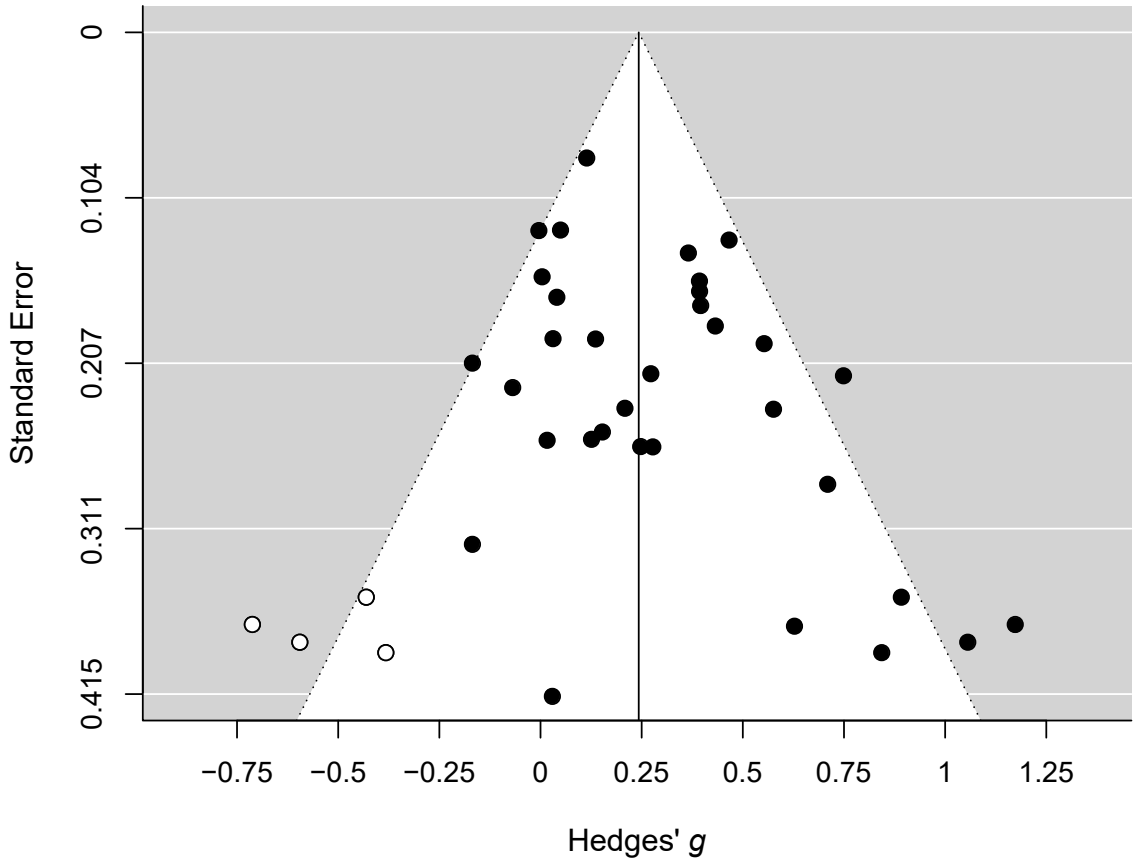

Figure 7. Funnel plot after Trim-and-Fill bias correction. Note that this analysis is based on the study-level effect sizes (Borenstein approach). Compared to the original funnel plot (see supplement), four studies were imputed to achieve symmetry (i.e., white circles). This resulted in a bias-corrected summary effect size of $g = 0.24$, CI_{95} [0.14, 0.34] that is slightly smaller than the original (Borenstein approach) estimate of $g = 0.28$, CI_{95} [0.19, 0.38].