OPEN ACCESS

University of Dundee

Reported self-control is not meaningfully associated with inhibition-related executive function

Saunders, Blair; Milyavskaya, Marina; Etz, Alexander; Randles, Daniel; Inzlicht, Michael

ORIGINAL RESEARCH REPORT

Reported Self-control is not Meaningfully Associated with Inhibition-related Executive Function: A Bayesian Analysis

Blair Saunders[*], Marina Milyavskaya[†], Alexander Etz[‡], Daniel Randles[§] and Michael Inzlicht[§,‖]

Self-control is assessed using a remarkable array of measures. In a series of five data-sets (overall N = 2,641) and a mini meta-analysis, we explored the association between canonical operationalisations of self-control: The Self-Control Scale and two measures of inhibition-related executive functioning (the Stroop and Flanker paradigms). Overall, Bayesian correlational analyses suggested little-to-no relationship between self-reported self-control and performance on the Stroop and Flanker tasks. The Bayesian meta-analytical summary of all five data-sets further favoured a null relationship between both types of measurement. These results suggest that the field's most widely used measure of self-reported self-control is uncorrelated with two of the most widely adopted executive functioning measures of self-control. Consequently, theoretical and practical conclusions drawn using one measure (e.g., the Self-Control Scale) cannot be generalised to findings using the other (e.g., the Stroop task). The lack of empirical correlation between measures of self-control do not invalidate either measure, but instead suggest that treatments of the construct of self-control need to pay greater attention to convergent validity among the many measures used to operationalize self-control.

Classically defined as the overriding of unwanted impulses (Baumeister, 2014; Roberts, Lejuez, Krueger, Richards, & Hill, 2014), self-control is among the most celebrated facets of higher cognition. Impulse control enjoys such theoretic prominence that inhibition has been proposed to underlie 80–90% of self-regulation (Baumeister, 2014; Baumeister, Heatherton, & Tice, 1994). Also remarkable is the array of measures used to assess self-control—from introspective self-report questionnaires to reaction-timed tests of executive functioning. While such measures are undoubtedly diverse, what seems to unite them is the idea that they each tap some ability to override unwanted dominant impulses (Baumeister et al., 2014; Baumeister, 2014; Hofmann, Schmeichel, & Baddeley, 2012; Inzlicht, Schmeichel, & Macrae, 2014).

That all these measures relate to a common construct called self-control has been assumed often by face-validity (e.g., "I am good at resisting temptations"; Stroop performance requires inhibiting word-reading),

but seldom investigated empirically (cf., Duckworth & Kern, 2011). Yet, the importance of understanding the nature of self-control measures cannot be understated. Poor self-regulation has been identified as "the major social pathology of the present time" (Baumeister et al., 1994, pp. 3; see also, Mischel, Shoda, & Rodriguez, 1989), with low reported self-control predicting poorer health, finances, and wellbeing, and higher rates of mortality and criminal convictions (Moffitt et al., 2011; Tangney, Baumeister, & Boone, 2004). Therefore, understanding operationalisations of the construct of self-control is critical. Most people have a lay conceptualisation of what it feels like to resist temptation and to exert self-control. However, are common empirical measures, that each putatively assess the ability to resist impulses in their own right, statistically related to each other? Do people who self-report high levels of self-control on questionnaire measures also show improved performance on laboratory tests of self-control?

Here, we addressed these questions by examining the relationship between self-report measurements of self-control and performance measures of inhibition-related executive functioning through five novel data-sets and a meta-analytical summary (N = 2,641).

The Self-Control Scale
Developed by Tangney and colleagues (2004), the Self-Control Scale is a self-report questionnaire assessing "the ability to *override* or change one's inner responses,

[*] Psychology, School of Social Sciences, University of Dundee, GB

[†] Department of Psychology, Carleton University, Ottawa, CA

[‡] Department of Cognitive Sciences, University of California, Irvine, US

[§] Department of Psychology, University of Toronto, CA

[‖] Rotman School of Management, University of Toronto, CA

Corresponding author: Blair Saunders (b.z.saunders@dundee.ac.uk)

as well as *interrupt* undesired behavioural tendencies (such as impulses) and *refrain* from acting on them" (pp. 274, emphasis added). Consistent with prominent theories (Baumeister et al., 1994), considerable emphasis was placed on inhibition in the development of the Self-Control Scale. Several items in the Self-Control Scale include content that is face-valid in relation to inhibition (e.g., "I am good at resisting temptation"; "I refuse things that are bad for me, even if they are fun"). Beyond inhibition, however, the Self-Control Scale assesses multiple outcomes and processes that are more globally reflective of self-discipline and goal-directed actions (e.g., "I have trouble concentrating"; "I tend to be disorganised"; "I say inappropriate things"; "I am lazy"; all reverse-scored).

Initial validation work demonstrated that higher Self-Control Scale scores were associated with better psychological adjustment, reduced problematic food and alcohol consumption, increased relationship satisfaction, and more adaptive emotional responses (Tangney et al., 2004). These relationships were largely confirmed in a subsequent meta-analysis of 102 studies (de Ridder, Lensvelt-Mulders, Finkenauer, Stok, & Baumeister, 2012). In short, self-control—as operationalised by the Self-Control Scale—predicts the good-life.

Likely reflecting the wide scope of the Self-Control Scale, this measure is strongly correlated with other broad individual differences including conscientiousness (John, & Srivastava, 1999; Roberts, Jackson, Fayard, Edmonds, & Meints, 2009) and grit (Duckworth et al., 2007). Conscientiousness encompasses needs for achievement, organization, restraint, and rule following (Goldberg, 1990; John, & Srivastava, 1999; Roberts, Jackson, Fayard, Edmonds, & Meints, 2009), and reliably predicts health, wealth, and well-being across the lifespan (Bogg & Roberts, 2004; Kern & Friedman, 2008). Similarly, grit is defined as perseverance and passion for long-term goals, even in the face of short-term setbacks and obstacles (Duckworth et al., 2007; Von Culin, Tsukayama, & Duckworth, 2014). Relative to self-control, conscientiousness and grit were not developed with a central emphasis on inhibition. However, given the strong correlations among grit, conscientiousness, and the self-control scale (cf., Roberts et al., 2014), our analyses also included a "self-discipline" composite that combined conscientiousness, grit, and the Self-Control Scale into a single measure.

Impulse control and executive functioning

In addition to self-report measurement, the ability to override impulses is commonly measured using executive functioning tasks such as the Stroop, Flanker, and Go-Nogo paradigms (Miyake et al., 2001). Bearing a close resemblance to definitions of self-control (Baumeister et al., 1994; Hofmann et al., 2012), executive functions reflect a domain-general range of processes that allow individuals to flexibly regulate attention and behaviour in a goal-directed manner (Banich, 2009; Miyake et al., 2000). Assessed using a range of tasks, executive functions demonstrate both unity and diversity. While a common factor appears to unite diverse measures of executive functioning, the processes of inhibition, updating, and

switching have also been identified as dissociable sub-components (Miyake et al., 2000). More recent analyses have suggested, however, that the inhibition component is particularly strongly related to the shared variance among executive functions (Miyake & Friedman, 2012). This latter finding is consistent with the idea that inhibition-related processes play a central role in governing flexible goal-directed actions.

In the present studies, we used the Stroop and flanker tasks as performance measures of inhibition-related executive functioning. In the Stroop task, people identify the physical colour of a word while ignoring the lexical meaning of the word that may be either compatible (the word "blue" in blue ink) or incompatible (the word "red" written in blue ink). Participants in the flanker task must respond to a central target letter (e.g., "S" or "H") while surrounding flanker letters prime either the correct response on compatible trials (e.g., "SSSSS"), or the incorrect response on incompatible trials (e.g., "HHSHH"). In both cases control is required to overcome the dominant but unwanted response tendency primed by the flankers or automatic word-reading processes on incompatible trials.

The Stroop and flanker tasks were originally conceived to investigate cognitive interference, and were presented in within-subjects experimental designs (Eriksen & Eriksen, 1974; Stroop, 1935). However, these paradigms have been widely adopted as self-control measures in social and personality psychology (Allom et al., 2016; Edmonds et al., 2009; Gailliott et al., 2007; Inzlicht & Gutsell, 2007; Molden et al., 2012). Despite their experimental origins, performance measures of inhibition-related executive functioning are frequently used in cross-sectional designs to assess individual differences in inhibition and impulsivity (Davidson, Amso, Anderson, & Diamond, 2006; Hall, 2012; Nigg, 2001; Snyder, 2012). Individual differences in executive functions have been related to multiple real-world outcomes thought to characterise good self-control, such as relationship fidelity (Pronk, Karremans, & Wigboldus, 2011), treatment-compliance among drug users (Streeter et al., 2008), and not snacking on unhealthy foods (e.g., Allan, Johnston, & Campbell, 2010). Thus, while tasks like the Stroop are often used to assess the influence of short-term, state effects on self-control (e.g., Inzlicht & Gutsell, 2007; Molden et al., 2012), it is variation in these tasks as an individual difference that is of interest in the current work. Whenever self-control on the Stroop task or flanker task is mentioned in this manuscript as a correlate of other measures (e.g., the Self-Control Scale), we are referring to self-control as an individual difference rather than a momentary, state effect.

One controversy regarding conflict control tasks is the specific role of inhibition in successful task performance. For example, interference from incompatible trials could be overcome either by inhibiting the inappropriate motor response to the irrelevant stimulus dimension (e.g., inhibiting the flanker letters; overriding word-reading processes in the Stroop task) or by focusing attention on the task-relevant dimension (e.g., the central letter in

the flanker stimulus, or the physical colour of the Stroop target; Cohen, Dunbar, McClelland, 1990; Egner & Hirsch, 2005). No clear consensus has emerged on whether these executive functioning tasks rely on selective attention, inhibition, or both. Most important for present concerns, however, is that these tasks more broadly assess the ability to control the influence of inappropriate impulses on performance in a goal-directed manner. Hereafter, these processes will be referred to as *inhibition-related executive functions*, while acknowledging that the exact mechanisms underlying successful performance on these tasks requires clarification.

The convergence between self-reported self-control and performance measures of self-control

The heterogeneity of tools available to investigate self-control might suggest that a single latent process underlies impulse control across contexts and modalities—from reaction time performance at a millisecond resolution to global introspections about one's ability to self-regulate.

This consilient view of self-control assessment is hindered, however, by the modest correlations among measures of self-control. A recent meta-analysis (Duckworth & Kern, 2011) reported small but statistically significant convergence among measures of self-control (i.e., self-report, informant report, choice tasks, and inhibition-related executive functioning, $r = .27$). This meta-analysis further revealed that while the expected correlations between self-reported self-control and inhibition-related executive functions were present, the effect size of these associations were particularly small in magnitude ($r = .10$).

Multiple factors might contribute to low convergence among measures of self-control. As mentioned previously, the self-control scale is not solely focused on inhibition, but incorporates a wide range of characteristics and behaviours that are indicative of self-discipline. The breadth of the Self-Control Scale might mean that it assesses a broad spectrum of characteristics, with this content only overlapping to a small extent with the inhibitory processes that are assessed in tests of inhibition-related executive functioning. It is also true that inhibition-related executive functions focus more narrowly on a constrained range of cognitive processes that allow people to overcome interference that is specific to a single, rather abstract, goal (e.g., "name colours", "identify the central target letter"). As such, this relatively narrow scope of the Stroop and flanker tasks might mean that inhibition-related executive functions can only correlate very slightly with the broader definition of inhibition used in the Self-Control Scale, simply because these executive functioning measures capture significantly less content. It is noteworthy that this logic still anticipates a detectable (albeit small) relationship between the broader trait measure and the performance measure.

The reliability paradox is another factor that might contribute to the relatively low correlations (Hedge, Powell, & Sumner, 2017). This paradox refers to a phenomenon where behavioural tasks only become established in the experimental literature if they show little between-subject variability. Consequently, low between-subject variability might limit the extent to which behavioural measures of self-control (e.g., Stroop, flanker) can correlate with other individual differences, such as the Self-Control Scale. Again, this reliability paradox might explain why correlations between self-reported self-control and inhibition-related executive functions would be rather small.

Finally, common biases in academic publishing might mean that prior meta-analytical effect sizes might overestimate the strength of relationship between scale measures of self-control and performance measures, like the Stroop or flanker tasks. The modest correlations between scale and performance measures reported in prior meta-analyses (i.e., Duckworth & Kern, 2011) might overestimate the underlying effect sizes because they did not correct for publication bias (i.e., the general tendency for significant results to be published more often than null results; Rothstein, Sutton, & Borenstein, 2005).

The current work

In a series of 5 data-sets and a meta-analysis, we asked if the Self-Control Scale (Tangney et al., 2004)—likely the most common self-control questionnaire—is correlated with the overriding processes commonly identified by the Stroop and flanker tasks. When formatting our research question, it is important to consider the smallest effect size that would still be of practical and theoretical interest. A correlation of at least moderate strength (e.g., $r \geq .4$) might be expected if moderate-to-strong convergence is anticipated between inhibitory executive functions and trait-self-control. However, effect sizes in this range seem particularly unlikely given prior meta-analytical results (Duckworth & Kern, 2011). As such, we anticipated that any correlation between scale and inhibition-related executive functioning measures of self-control would be small (e.g., $rs > .10 < .25$). It should be further noted that the smallest effect size of theoretical significance differs depending on the proposed application, however, it seems that any effect size approaching zero ($<.10$) would be small enough to be off little practical or theoretical significance.

Our work builds on prior investigations in two key ways. First, working against potential file-drawer problems, we present every study that we are aware of from our laboratory (i.e., the Toronto Laboratory for Social Neuroscience) that includes the Stroop or flanker task and the Self-Control Scale. Second, by exclusively employing null-hypothesis significance testing, existing studies can only support convergence among measures. Instead, we used Bayesian methods to obtain a Bayes factor comparing a null and non-null hypothesis, and computed the posterior distribution of the correlation given the alternative model. The Bayes factor gives us the evidence for or against the hypothesis of no correlation, and the posterior tells us how big any non-zero correlation is likely to be (see Etz & Vandekerckhove, in press).

Study 1

The Stroop task is a common laboratory operationalisation of self-control (e.g., Gaillott et al., 2007; Inzlicht & Gutsell, 2007; Molden et al., 2012). In study 1, we used the Stroop

task to test convergence between reported self-control and inhibition-related executive functioning.

Method
Participants
We decided apriori to collect data from at least 200 participants. This sample size exceeds that of similar prior investigations (Allom et al., 2016; Edmonds et al., 2009), and meets rule-of-thumb guidelines for sample size in social-personality psychology (Fraley & Vazire, 2014). 224 undergraduate students from the University of Toronto Scarborough participated for course credit (81 females; mean age = 18.8, *SD* = 2.5 years). Seven participants were excluded from these analyses either because of software malfunction or incorrectly responding to likert-type questions (responding outside the range of the scale).

For each study in this manuscript, ethical approval was obtained from the Research Ethics Board at the University of Toronto, Scarborough before data collection started. In each study informed consent was provided by each participant before taking part.

Scales
Self-control was assessed using the 13-item Self-Control Scale (Tangney et al., 2004). We additionally measured two other self-regulatory individual differences: conscientiousness and grit. Conscientiousness was assessed using the conscientiousness 9-item subscale of the 44-item Big Five Inventory (John & Srivastava, 1999). Grit was assessed using the 12-item grit scale (Duckworth et al., 2007). Please see **Table 1** for descriptive and reliability statistics for all scales. Although theoretical distinctions exist between grit, conscientiousness, and the Self-Control Scale (cf., Duckworth & Gross, 2014), these self-regulatory traits tend to correlate highly with each other and the Self-Control Scale (*r*s ~ .70–.80; Roberts et al., 2013). Consequently, we created a composite self-discipline measure aggregating across the Self-Control Scale, conscientiousness, and grit. Similar composite scores have previously demonstrated higher utility in predicting the ability to select between competing impulses than single measures (Duckworth & Seligman, 2006). Thus, by improving signal-to-noise ratios, this composite scale might show a particularly reliable relationship with inhibition-related executive functioning. Scales were computerized and completed by participants immediately after the Stroop task.

Stroop paradigm
The Stroop started with 10 practice trials of 6 object-words (chair, house, lamp, spoon, table, and window) presented in red or blue. The left arrow-key was pressed to identify blue words, and the right arrow key for red words.

The main Stroop task started after these practice trials. Stimuli were the words "RED" and "BLUE" presented in either red or blue to create compatible (e.g., "RED" in red font) and incompatible (e.g., "BLUE" in red font) targets. Trials commenced with a fixation cross (250 ms), followed by a target stimulus until response (min: 150 ms; max:

1500 ms) followed by a blank screen (550 ms).[1] Self-paced breaks occurred between blocks where participants reported their subjective experience (not analysed here).

We created mean scores for each dependent measure (Stroop effects [i.e., RT/error-rates on incompatible trials *minus* RT/error rates on compatible trials], accuracy and reaction times on compatible trials and incompatible trials) separately for each experimental block to assess the internal consistency of the Stroop task. Consequently, Cronbach's α was calculated using 9 different values (i.e., one from each experimental block) for each measure in the experiment (e.g., the Stroop effect in RT). The resulting internal consistency estimates suggested that reliabilities were good-to-excellent for most measures (compatible mean RT, α = .95; incompatible mean RT, α = .93; compatible error rates, α = .83; incompatible error rates, α = .83), but low for the Stroop effect in mean reaction time (α = .52) and error rates (α = .46; see also Wöstmann et al., 2013).

Given these reliabilities, it might be suggested that the subsequently reported null correlations were a result of low measurement reliability, rather than low association among reported self-control and inhibition-related executive function. However, it should be noted that similar small-to-null correlations were observed when Bayesian correlations were conducted between reported self-control/self-discipline and non-difference executive functioning scores (i.e., mean RT and error-rates for incompatible and compatible trials) across all studies in this manuscript (see supplemental materials).

Analysis strategy
Bayesian Pearson's correlations were computed to quantify the association between the Stroop effect and self-report scores. The primary benefit of adopting this Bayesian approach is that in addition to providing evidence in favour of an alternative hypothesis (e.g., a negative correlation between the Self-Control Scale and the Stroop effect) and estimating the size of the correlation, Bayes factors can also be used to provide evidence in favour of a null relationship (Wagenmakers, Verhagen, & Ly, 2015). Bayes factors can be interpreted "as the degree to which the data sway our belief from one to the other hypothesis" (Etz & Vandekerckhove, 2016, p. 4). For example, in the case where we initially have no preferred hypothesis, so that we assign 50% prior probability for each of the null and alternative hypotheses, a Bayes factor of 3 in favor of the null brings the probability of the null hypothesis to 75% (a Bayes factor of 10 brings the probability of the null hypothesis to 91%).

Results and Discussion
Correlations among self-regulatory constructs
As anticipated, grit, conscientiousness, and Self-Control Scale scores were strongly correlated (all *r*s > .598, see **Table 1**). A self-discipline composite measure was created by z-scoring grit, the Self-Control Scale, and conscientiousness and summing these standardised values.

Table 1: Table showing correlations (Pearson's *rho*) and descriptive statistics for the Self-Control Scale, Conscientiousness, and Grit.

		Scale				Descriptives	
		Self-disc.	Self-control	Grit	Consc.	M (SD)	α
Study 1	Consc.	.892	0.695	0.644	–	3.32(0.69)	.75
	Grit	.855	0.598	–		2.96(0.67)	.83
	Self-Control	.875	–			3.19(0.57)	.77
	Self-disc.	–				0.00(2.6)	.85

Note: Consc. = Big Five Inventory – Conscientiousness; Self-Control = Self-Control Scale; Self-disc = Self-Discipline Composite; α = Cronbach's alpha.

Stroop effects

Our procedure successfully produced the Stroop effect in reaction time and error rates. Responses were faster on compatible ($M = 438$ ms; $SE = 3.80$) than incompatible trials ($M = 455$ ms; $SE = 4.42$) trials, $t(216) = −12.579$, $p < .001$, $d = −0.947$. Error rates were higher on incompatible ($M = 7.5\%$; $SE = 0.35$) than compatible trials ($M = 4.9\%$ $SE = 0.28$), $t(216) = −11.521$, $p < .001$, $d = −0.821$.

Bayesian correlations

We next tested correlations between the Stroop effect and the self-report measures. Difference values were used because they are indicative of overriding processes while controlling for base rate performance.[2] As increasing values on each self-report scale should be associated with reduced Stroop effects (i.e., higher control), we set a prior on the correlation that is skewed such that most of its mass (78%) falls on negative correlation values (see **Figure 1**).[3] Here, a Bayes factor in favour of the alternative hypothesis ($BF_{01} < 1$) would support correlations among performance and self-report measures. In contrast, a Bayes factor favouring the null ($BF_{01} > 1$) would suggest no relationship between questionnaire and behavioural measures. We report posterior medians and 95% posterior (credible) intervals in brackets for each correlation, which tell us the most likely range for the value of the correlation if it is in fact non-zero.

Stroop effect in reaction time. The data provide evidence that the Stroop effect in reaction times was not correlated with Self-Control Scale scores ($r = −.012$ [−.143, .119] $BF_{01} = 7.73$) or the self-discipline composite measure ($r = −.027$ [−.158, .105], $BF_{01} = 7.26$), see **Figure 2**. Moreover, the posterior intervals suggest that any correlation between reaction time performance and these scales would likely be small.

Stroop effect in choice error rates. The data provide some evidence that the Stroop effect in error rates was not associated with scores on the Self-Control Scale ($r = −.067$ [−.196, .065], $BF_{01} = 4.81$) or the self-discipline composite measure ($r = −.102$ [−.231, .029], $BF_{01} = 2.44$), see **Figure 2**. The posterior intervals suggest that any correlation between error rates and either of these scales would most likely be negative and small.[4]

Studies 2 a-c

The following three studies replicated the above findings using a large online sample of participants and another common test of inhibition-related executive functions, the Flanker paradigm (cf., Eriksen & Eriksen, 1974; Miyake et al., 2001). Because these studies share an almost identical protocol we present them together noting any methodological divergence.

### Method
Participants

Study 2a. Eight hundred and fifty-six participants (397 females; mean age = 35.23, $SD = 11.11$ years) were recruited from Mechanical Turk (MTurk) in October 2015. MTurk is crowd-sourced marketplace in which workers are remunerated for completing online tasks. While data from MTurk tends to be noisier than lab data, this platform facilitates the recruitment of lager samples in service of statistical power (Crump, McDonnell, & Gureckis, 2013). All recruited MTurk workers were located in the USA and had completed a minimum of 5 HITs with >90% success rate. Participants were compensated $3 USD, and sessions lasted ~25 minutes. The study consisted of an arrow version of the flanker task, followed by a battery of self-report questionnaires. This data was initially collected as a training sample for a machine learning project. However, we only analysed variables pertinent to the hypotheses addressed in Study 1: The flanker data; reported self-control, conscientiousness, and grit.

Two hundred and nine participants were excluded for making >40% errors on the flanker task or having missing data from the flanker task (e.g., because of software malfunction).[5] While this exclusion rate is high (>24%), the 40% error criteria ensures that responding is above chance. One further participant was excluded for very fast mean reaction time (<100 ms). The final sample included 647 participants, resulting in 80% power to find a correlation of .11 when $\alpha = .05$; and >99% power to find a correlation of .20.

Study 2b. 1,621 participants (1227 females, mean age =39.20, $SD = 14.22$) recruited by the online survey and market research company Tellwut (www.tellwut. com). As with the previous sample, all participants were recruited from the USA, paid $3 USD for taking part, and completed the same study protocol as in
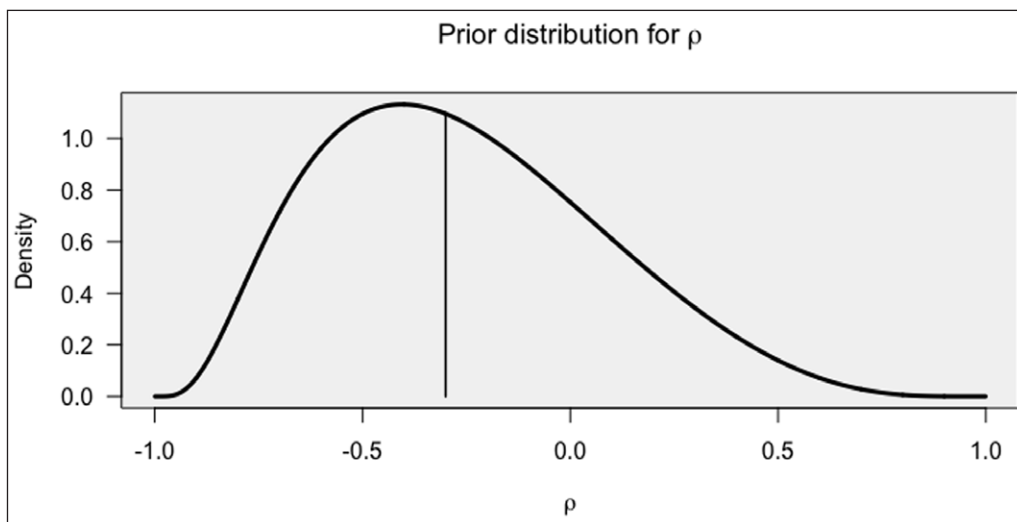
Figure 1: Prior distribution used in the Bayesian correlation analysis in Study 1. The vertical line marks the median of this distribution at –.29. This prior was obtained by first placing a N(–.3, .4) prior on Fisher's Z and then converting back to the correlation scale (see the supplementary materials for details, https://osf.io/8etus/). Correlation coefficients unaffected by the prior can be found in the online code on our OSF page.
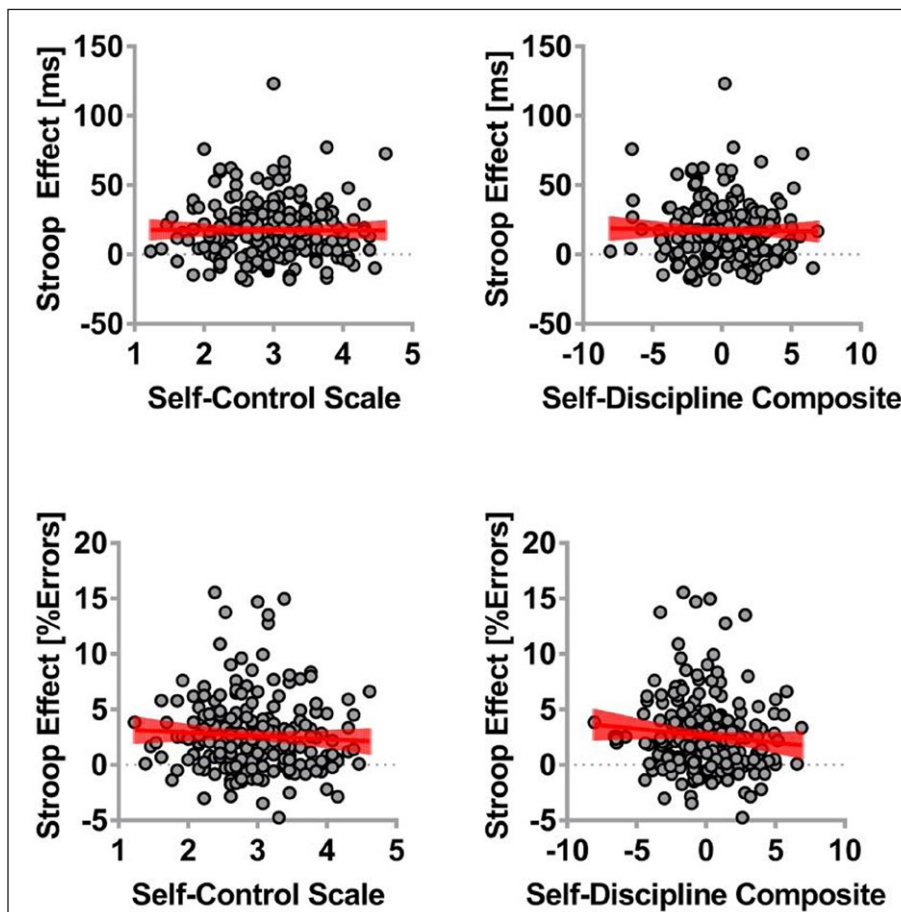
Figure 2: Scatterplots show the Stroop effect in reaction time (top panels) and error rates (lower panels) as a function of the Self-Control Scale (left) and the self-discipline composite measure (right). Lines of best fit show linear regressions including 95% confidence intervals.

Study 2a. Exclusion criteria were identical to study 2a resulting in 900[6] participants being excluded and a final sample of 723 participants. Upon visual inspection of the scatterplots one participant was also removed whose flanker effect that was >4 SDs below the mean.

Study 2c. 1,769 participants (932 females, mean age = 40.21, *SD* = 12.53) were recruited from the online panel company Cint (www.cint.com). These participants were also based in the USA, were paid $3 USD for participation, and completed a protocol that was identical to studies 2a and 2b. The same exclusion criteria applied to studies 2a and 2b identified 888 participants for exclusion resulting in a final sample of 881 individuals.

Scales
In each study, grit and conscientiousness were assessed as in Study 1, while we used the extended 36 item version of the Self-Control Scale (cf., Tangney et al., 2004). See **Table 2** for descriptive statistics. Conscientiousness was assessed in study 2c using the relevant subscale from the IPIP-NEO-120 (Johnson, 2014).

Arrow Flanker task
All three studies used an identical arrow flanker task. Flanker stimuli consisted of five arrowheads (i.e., compatible: <<<<<, >>>>>; incompatible: <<><<, >><>>). Participants responded to the central arrow using the left or right arrow-key on their keypad. Here, control is required on incompatible trials to override the misinformation primed by the flankers. The main experiment consisted of 100 flanker trials (50 compatible, 50 incompatible) preceded by 10 practice trials. Each trial started with a fixation cross for 1000 ms, followed by the target until response (max. 1000 ms).

Our statistical approach followed that used in Study 1. Due to the short nature of the flanker task in these studies, we were unable to obtain comparable measures of internal consistency in the brief online flanker task as we did in Study 1 for the Stroop task. However, previous reports have suggested that the reliability of the Flanker is similar to that which we obtained in Study 1 (Wöstmann et al., 2013).

Results and Discussion
Scale correlations and reliability
All three studies supported strong positive associations among self-reported grit, self-control, and conscientiousness, all *r*s > .698, all *p*s < .001, see **Table 2**. A self-discipline composite measure was created as in Study 1.

Flanker effects
These brief online experiments revealed robust Flanker effects in all three data-sets: Responses were slower and more error-prone on incompatible than compatible trials (see **Table 3**).

Bayesian correlations
To evaluate the correlations from studies 2a, 2b, and 2c we employed the following strategy: For each analysis in study 2a, we used the posterior from the corresponding test in study 1 as the prior. Then for each analysis in study 2b, we used the posterior from 2a as the prior, and repeated this again for 2c. This Bayesian updating yields sequential Bayes factors (BF_{0r}; see Ly, Etz, Marsman, & Wagenmakers, 2017), and these can be interpreted as the additional evidence gained from each study beyond the evidence we had before.

Flanker effect in mean reaction times. The data from studies 2a and 2b provide little additional evidence for discerning whether scores on the Self-Control Scale are associated with the flanker effect or not (2a: *r* = .013 [−.054, .080], BF_{0r} = 1.87; 2b: *r* = −.009 [−.058, .040], BF_{0r} = 1.37), whereas 2c provides some counter evidence in

Table 2: Table showing correlations (Pearson's *rho*) and descriptive statistics for the Self-Control Scale, Conscientiousness, and Grit.

		Scale				Descriptives	
		Self-disc.	Self-Control	Grit	Consc.	M(SD)	α
Study 2a	Consc.	.931	.792	.790	–	3.89(0.74)	.895
	Grit	.921	.764	–		3.51(0.74)	.897
	Self-Control	.922	–			3.62(0.62)	.936
	Self-disc.	–				0.01(2.77)	.915
Study 2b	Consc.	.897	.700	.711	–	3.74(0.74)	.849
	Grit	.9	.699	–		3.38(0.62)	.791
	Self-Control	.894	–			3.36(0.59)	.905
	Self-disc.	–				0.01(2.66)	.879
Study 2c	Consc.	.913	.775	.726	–	5.17(0.89)	.892
	Grit	.905	.745	–		3.42(0.57)	.765
	Self-Control	.924	–			3.48(0.58)	.913
	Self-disc.	–				0.001(2.74)	.901

Note: Consc. = Big Five Inventory – Conscientiousness; Self-Control = Self-Control Scale; Self-disc. = Self-Discipline Composite; α = Cronbach's alpha.

Table 3: Table showing descriptive and inferential statistics for the Flanker tasks for studies 2a–c.

		Flanker effects			
		Compatible	Incompatible		
		M(SD)	M(SD)	*t*	Cohen's *d*
Study 2a	RT (ms)	452(58)	474(61)	−14.940***	−0.563
	%Errors	5.25(6.27)	8.19(9.74)	−9.159***	−0.389
Study 2b	RT (ms)	456(56)	481(56)	−20.406***	−0.759
	%Errors	9.09(8.74)	13.70(12.97)	−12.114***	−0.486
Study 2c	RT (ms)	449(61)	478(60)	−20.916***	−0.995
	%Errors	9.00(8.60)	15.01(14.83)	−14.450***	−0.544

Note: RT = mean reaction time; %Errors = mean choice error rate; *** = $p < .001$.

favour of a non-zero correlation (2c: $r = −.045$ [−.084, −.006], $BF_{0r} = .107$, see **Figure 3**. A similar pattern holds for the associations between the flanker effect and Self-discipline composite scores (2a: $r = .024$ [−.043, .090] $BF_{0r} = 1.68$; 2b: $r = .005$ [−.044, .054], $BF_{0r} = .1.69$; 2c: $r = −.026$ [−.065, .014], $BF_{0r} = .568$), see **Figure 4**.

Choice error rates. The data from studies 2a–2c provide mixed evidence for whether flanker effect in error rates are associated with the Self-Control Scale (2a: $r = −.009$ [−.076, .058], $BF_{0r} = 3.12$; 2b: r = −.035 [−.084, .014] $BF_{0r} = .535$; 2c: $r = −.061$ [−.010, .021], $BF_{0r} = .035$). A similar overall pattern holds for the association between flanker effect error rates and the self-discipline composite (2a: $r = −.014$ [−.081, .052], $BF_{0r} = 5.81$; 2b: $r = −.024$ [−.073, .026], $BF_{0r} = 0.949$; 2c: $r = −.045$ [−.085, −.006], $BF_{0r} = .162$).

Study 3
Study 3 was an in-lab version of the flanker task that was conducted as part of a battery of baseline measures for a longitudinal goal-striving study that was on-going at the time of writing this manuscript. The flanker task always came first in a series of 3 tasks that were undertaken by participants while electroencephalography (EEG) was recorded. The other two tasks in this series were a time-estimation paradigm and passive picture viewing task that were not relevant to the current research questions. After these tasks the participants answered a battery of questionnaires that included assessments of conscientiousness, grit, and trait self-control. The following analyses report only on the executive functioning and self-report measures included in studies 1 and 2a–c.

Method
Participants
Two hundred and twenty-one participants took part in the study (61.5% females, mean age = 20.2, SD = 5.6), and were largely recruited through the undergraduate pool at the University of Toronto Scarborough, while a smaller number of participants were recruited through community advertisements. The session lasted approximately two hours, including setting up the EEG apparatus, completing the computerised tasks, and the subsequent questionnaires. Forty-three participants were excluded following the same exclusion criteria that were applied in Studies 2a–c.

Scales
Self-control and grit were assessed in the same manner as Study 1 (see **Table 4**), however, the likert questions for the Self-Control Scale were a range of 7 points rather than 5. Grit was assessed by the 8-item short grit scale (Duckworth & Quinn, 2009; $\alpha = .76$). The questionnaires were answered by participants after the computerised tasks.

Arrow-flanker task
As in studies 2a–c, participants performed an arrow version of the flanker task in a dimly lit room. Here, we only note deviations from the protocol used in these previous studies. Trials commenced with the presentation of a fixation cross (250 ms) that was followed by a flanker target stimulus until response (min: 100 ms, max: 1000 ms), followed by a blank screen for 600–1000 ms before the start of the next trial. Participants performed a total of 420 trials. Participants were given self-paced breaks after blocks of 60 trials, and were instructed to respond as quickly and accurately as possible.

Results and Discussion
As with all previous studies, strong positive correlations were observed between conscientiousness, trait self-control, and grit, all $rs > .676$, $ps < .001$.

Flanker effects
Responses were slower on incompatible trials ($M = 488$ ms, $SD = 44.24$) than compatible trials ($M = 400$ ms, $SD = 34.27$), $t(177) = −46.08$, $p < .001$, $d = −3.481$. Error rates were also higher for incompatible trials ($M = 18.9\%$, $SD = 12.55$) than compatible trials ($M = 2.3\%$, $SD = 4.16$), $t(177) = −20.07$, $p < .001$, $d = −1.504$. Together, these results confirm that the classic flanker interference effect was present in this data set. As in the previous studies, in
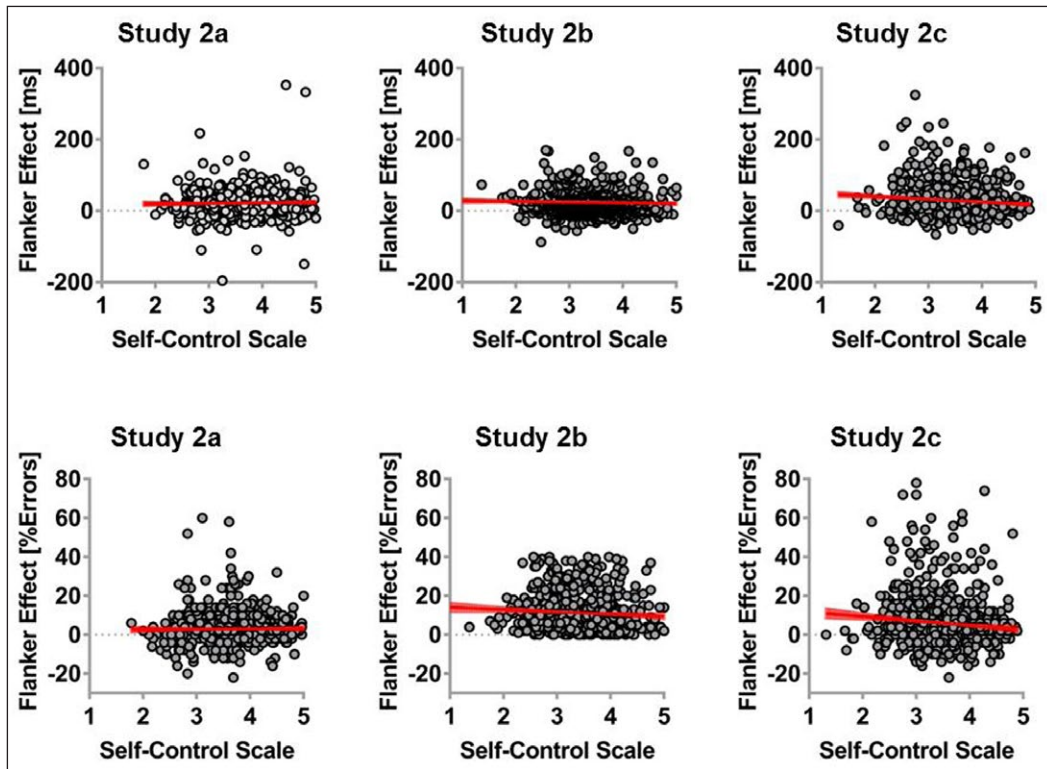
Figure 3: Scatterplots show the Flanker effect in reaction time (top panels) and error rates (lower panels) as a function of the Self-Control Scale. Lines of best fit show linear regressions including 95% confidence intervals.
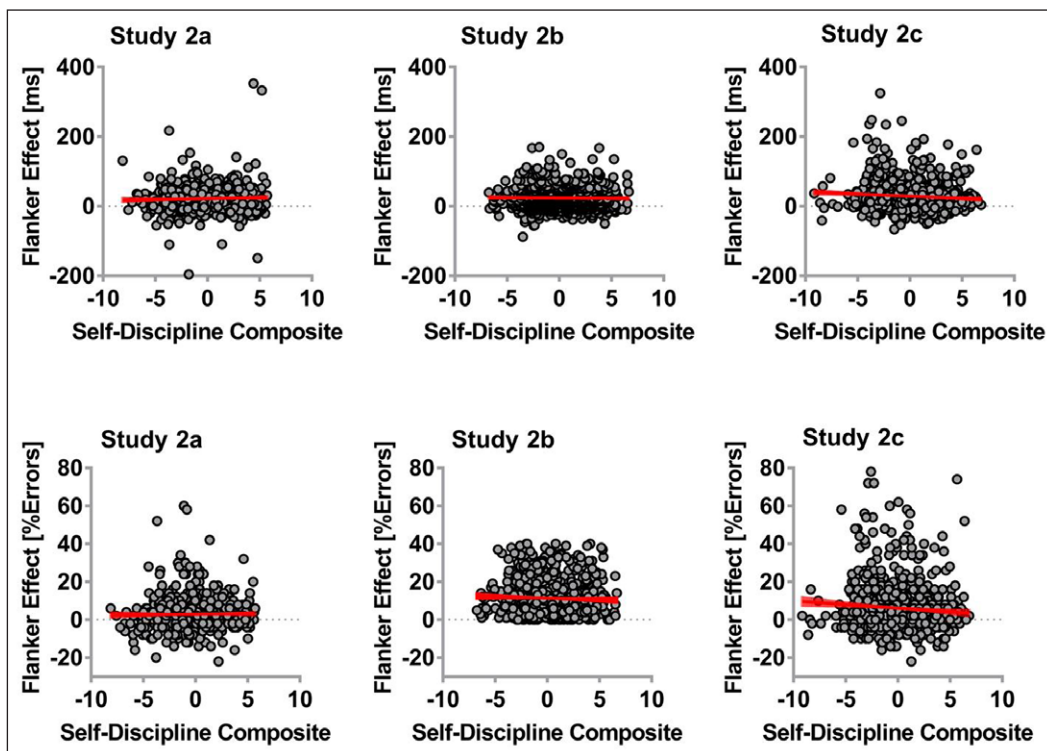
Figure 4: Scatterplots show the Flanker effect in reaction time (top panels) and error rates (lower panels) as a function of the self-discipline composite measure. Lines of best fit show linear regressions including 95% confidence intervals.

Study 3 we used the difference scores (incompatible *minus* compatible) created from the flanker task separately for reaction times and error-rates.

Bayesian correlations

Following the strategy of Bayesian updating from the previous studies, the posterior from Study 2c was used as

Table 4: Table showing correlations (Pearson's *rho*) and descriptive statistics for the Self-Control Scale, Conscientiousness, and Grit.

		Scale				Descriptives	
		Self-disc.	Self-control	Grit	Consc.	M(SD)	α
Study 3	Consc.	.893	.697	.682	–	3.37(0.64)	.76
	Grit	.884	.676	–		3.26(0.62)	.83
	Self-Control	.892	–			4.06(1.00)	.76
	Self-disc.	–				0.01(0.89)	.92

Note: Consc. = Big Five Inventory – Conscientiousness; Self-Control = Self-Control Scale; Self-disc. = Self-Discipline Composite; α = Cronbach's alpha.

the prior for study 3 to yield a further sequential Bayes factor (Ly et al., 2017). We can interpret each posterior from study 3 as the result of a Bayesian fixed-effects meta-analysis because they contain all the information accumulated across all the studies. Furthermore, a fixed effect meta-analytic Bayes factor is given by the product of all individual studies' sequential Bayes factors. For mean reaction times, the data from Study 3 provided a little more evidence in support of a null relationship between the self-control scale and the flanker effect, $r = -.039$ $[-.077, 0]$, $BF_{0r} = 1.8$, with the overall evidence favouring no association (fixed effect meta-analysis $BF_{01} = 3.81$). A similar result was obtained for the flanker effect in reaction time for the self-discipline composite, $r = -.024$ $[-.062, .015]$, $BF_{0r} = 1.12$, with the overall evidence again favoring no association (fixed effect meta-analysis $BF_{01} = 13.1$). The data from the flanker effect in error rates provided a little more evidence favoring a small association with the self-control scale, $r = -.060$ $[-.098, -.022]$, $BF_{0r} = .848$, with the overall evidence favoring a small association (fixed effects meta-analysis $BF_{01} = .239$). The association between error rates and the self-discipline composite showed a similar pattern in this study, $r = -.046$ $[-.085, -.008]$, $BF_{0r} = .763$, but the overall evidence was essentially equivocal (fixed effects meta-analysis $BF_{01} = 1.67$).[7]

Bayesian Random Effects Meta-Analysis
We found converging lines of evidence for zero to small correlations between inhibition-related executive functions and reported self-control across five data-sets. We next conducted a Bayesian random-effects meta-analysis to estimate the overall size of the correlations while accounting for potential heterogeneity of results across the data sets. To this end, four separate meta-analyses (reaction time and choice-error rates separately for the Self-Control Scale and self-discipline composite) were conducted using the Bayesian statistical software Stan (Carpenter, et al., 2016; Stan Development Team, 2017), and we computed meta-analytic Bayes factors using bridge sampling (Gronau et al., 2017; Gronau, Singmann, & Wagenmakers, 2017). These random effect meta-analyses were instantiated as hierarchical models where the individual studies can be related to each other through shared population-level parameters; see the supplementary materials for details.

The results of each random-effects meta-analysis are summarized in **Figure 5**. The posterior distribution for the meta-analytic correlation (rho) between the Self-Control Scale predicting conflict effects in reaction time suggests any association that might exist is likely negative and small ($r = -.028$ $[-.181, .109]$), and the Bayes factor favours the null model ($BF_{01} = 8.34$). Similarly, the posterior for the meta-analytic correlation between the self-discipline composite score and conflict effects in reaction time was small ($r = -.021$ $[-.161, .093]$), and again the Bayes factor favoured the null model ($BF_{01} = 10.33$).

A similar pattern of results holds for the meta-analytic correlations between the conflict effects on error rates and scores the Self-Control Scale ($r = -.059$ $[-.190, .056]$), with a Bayes factor slightly favouring the null model ($BF_{01} = 3.93$). Similar results were found for the self-discipline composite measure ($r = -.051$ $[-.183, .057]$, $BF_{01} = 4.87$).

General Discussion
Combining 5 data-sets with over 2,600 participants, we found a consistent pattern of a small-to-zero relationship between self-reported self-control and two performance measures of inhibition-related executive functioning (the Stroop and flanker tasks). Most individual studies were consistent with no association between self-reported self-control and inhibition-related executive functions, with only study 2c tending to support a small negative relationship. Further Bayesian meta-analyses—both fixed and random effect—suggested little to no relationship between reported self-control and conflict effects in reaction time and choice error rates. This conclusion is supported by both Bayes factors supporting a null association and the corresponding small posterior estimates.

What do these results mean for the science of self-control?
Both questionnaire and inhibition-related performance measures are established and widely accepted measures of self-control (de Ridder et al., 2012; Duckworth & Kern, 2011; Hofmann et al., 2012; Molden et al., 2012; Inzicht & Gutsell, 2007). Indeed, it was once estimated that inhibition underlies 80%–90% of self-regulation (Baumeister et al., 1994; Baumeister, 2014). Given that inhibition is evoked as a mechanism underlying executive functions (Miyake
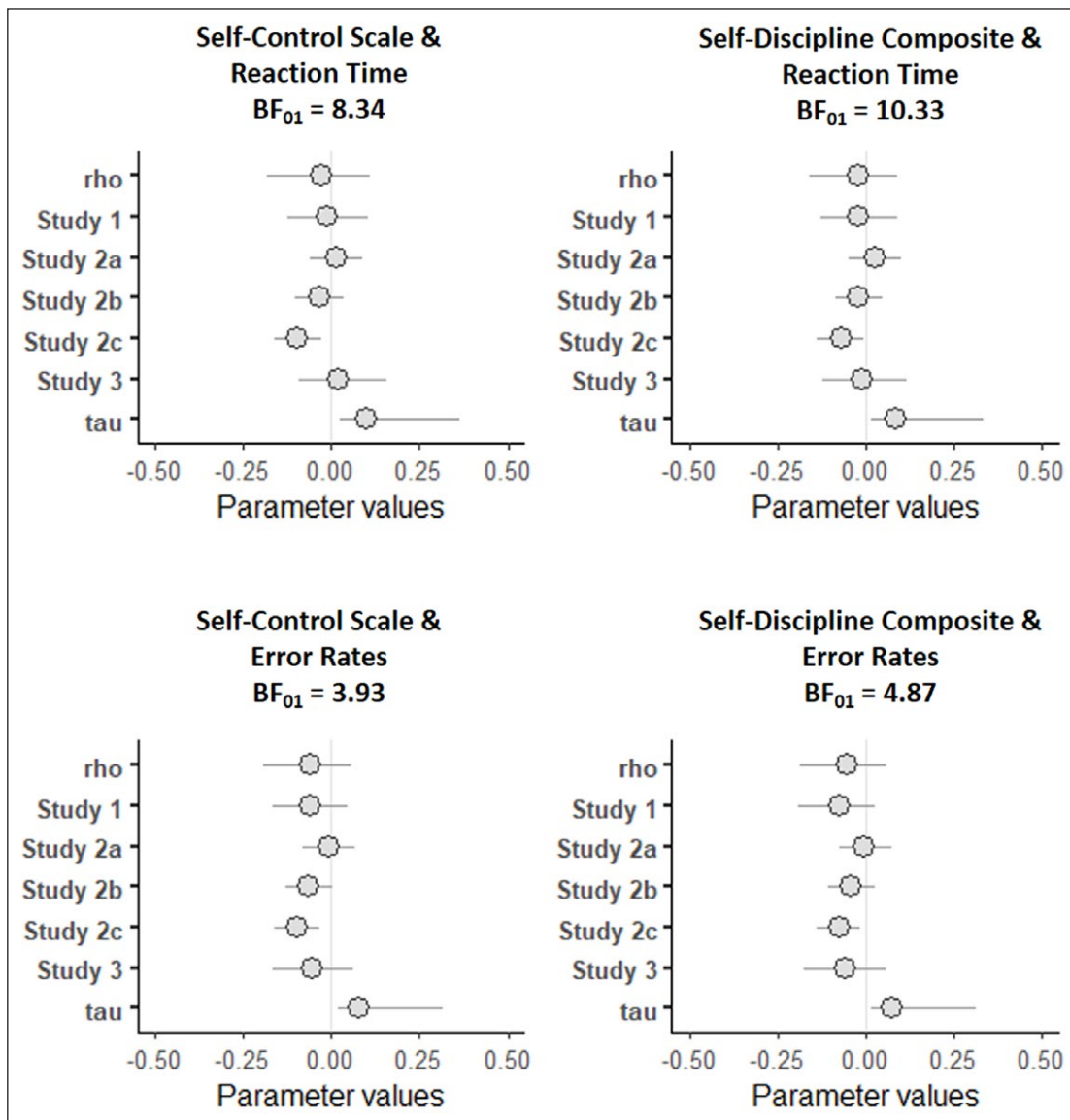
Figure 5: Forest plots depicting the results of the random effects meta-analyses as a function of self-report measure (Self-Control Scale, Self-Discipline Composite) and the difference scores on the executive functioning tasks for both reaction time and choice error rates. Error bars depict 95% confidence intervals. BF_{01} in sub-titles show evidence favouring the null for the meta-analytical (rho) effect across all four data-sets. Tau is a measure of heterogeneity of effect size, and was low for each of our meta-analyses.

& Friedman, 2012) and self-reported self-control (Tangney et al., 2004), it might be expected that associations among these self-control measures should be sizeable and robust. However, the current results suggest that questionnaire measures of self-control and canonical performance measures of inhibition-related executive function are largely unrelated to each other. Importantly, we do not claim that our results invalidate one measure or the other. Our results suggest that the Stroop and flanker tasks do not reflect the broader individual difference construct that is reflected in self-report scales, and, equally, that scores on the self-control scale are not analogous to the processes assessed by the Stroop and flanker tasks.

Our findings are consistent with previous studies that reported non-significant correlations between inhibition-related executive functions and self-report measures of impulsivity (Eisenberg et al., 2018; Nęcka, Gruszka,

Orzechowski, Nowak, & Wójcik, 2018; Stahl et al., 2013) or conscientiousness (Flemming, Heintzelman, & Bartholow, 2016). We further extend these frequentist analyses by providing Bayesian support for a null relationship between these measures. The strongest interpretations of these findings are: a) that theoretical and practical conclusions drawn using one measure (e.g., the Self-Control Scale) cannot be generalised to findings using the other (e.g., the Stroop task); and b) that there is little-to-no relationship among these measures that are both commonly identified as operationalisations of the psychological construct of self-control.

Executive functioning paradigms such as the Stroop and Flanker task are designed specifically to assess control over pre-potent impulses (cf., Botvinick et al., 2001; Miyake et al., 2001). Evidence that these tasks assess the ability to overcome inappropriate impulses has been

suggested both by behavioural and psychophysiological investigations (Kopp et al., 1996; Verleger et al., 2009). Similarly, the concept of inhibiting unwanted impulses was central to the development of the Self-Control Scale (Tangney et al., 2004), and many of the items in this scale assess inhibition-like content (e.g., "I am good at resisting temptation"). Despite these logical similarities, it is reasonable to conclude from our results that self-report measures of control and laboratory tests of inhibition-related executive functions assess different underlying processes. These findings should be of great concern to psychological scientists interested in self-control: Despite theoretical suggestions to the contrary (Hofmann et al., 2012), our results suggest that the field's most widely used trait measure of self-control is uncorrelated with two of the field's most commonly used executive functioning measures of self-control.

We should be clear that our results do not undermine the validity of the Self-Control Scale (or other self-report measures like it) as a predictor of real-world outcomes. Scale measures of self-control are consistently related to multiple indices of wellbeing (de Ridder et al., 2012; Moffitt et al., 2011). In fact, initial validation work focused on associations between the Self-Control Scale and relevant outcomes (e.g., less binge eating, alcohol abuse, better relationships, and good psychological adjustment), rather than exploring associations with other established self-control measures (Tangney et al., 2004). Instead, it appears that the Self-Control Scale and performance measures of inhibition-related executive functioning might be largely non-overlapping, despite these tasks both being framed as assessments of the ability to override impulses.

The current results are illustrative of a general conceptual and definitional ambiguity that may hinder the empirical validation self-control as a psychological construct. Self-control is typically defined as the ability to override unwanted impulses (Baumeister et al., 2014), and this ability is assessed using multiple measures. This heterogeneity of assessment is particularly challenging for construct validity as there currently exists no gold-standard criterion measure against which the validity of other self-control measures can be assessed. If I hypothesize that measure A (e.g., handgrip strength, the Stroop task, or a new self-report scale) is a measure of self-control, there is no agreed benchmark assessment of self-control that I can correlate with handgrip strength. Following this logic, it is impossible to conclude from our results if the observed lack of significant correlations points to validity issues with any one of our measures, or if there are broader problems with the construct space of self-control (cf., Cronbach & Meehl, 1955).

Conclusions complementary to our own were drawn in recent analyses in which task/performance based measures of self-regulation (e.g., go/no-go task, delay discounting, and 35 other behavioural tasks) predicted other task-based measures, but were not associated with 27 self-report based measures of self-regulation (and vice versa; Eisenberg et al., 2018). Together with our own results, these studies indicate a so-called *jingle-fallacy*

has emerged in self-control research, where two types of task (i.e., behavioural and self-report) that are commonly identified as operationalisations of one psychological construct (i.e., self-control), bear little-to-no empirical relationship with each other.

Limitations and future directions
The current studies should be considered in light of some important limitations and questions for future research. First, we focused on a relatively constrained range of self-report and performance measures that reflect canonical measures of self-reported self-control and inhibition-related executive functions. However, self-control and self-discipline can be measured both by longer and shorter-form scales (Goldberg, 1992; Gosling, Rentfrow, & Swann, 2003; Duckworth & Quinn, 2009), and also by observer reports (Jackson et al., 2010; Moffitt et al., 2011). Similarly, a wide range of reaction timed tasks are available to measure inhibition-related executive functions (e.g., antisaccade task, the Stop-signal task). Given this diversity of measures, it is clear that on-going research should test the generalisability of our findings to other measures in order to further explore the structure of self-control.

As mentioned when introducing our measures, there also exists a unity and diversity among established measures of executive functioning (Miyake et al., 2000; Miyake & Friedman, 2012). Future research could explore associations among self-report measures of self-control and other aspects of executive functioning, such as updating and task switching. One recent investigation indicated that the personality dimension of conscientiousness was associated with shifting, but not with the inhibition of prepotent responses or working memory updating (Fleming, Heintselman, & Bartholow, 2016). These findings suggest that rather than reflecting the ability to overcome impulses, conscientiousness might be more closely associated with control processes that allow people to flexibly respond to changing contexts and environments. Similar patterns might be expected with reported self-control given the high degree of empirical and conceptual overlap between self-control and conscientiousness (Roberts et al., 2014). Such a finding would be consistent with theoretical perspectives in which self-control involves adaptively managing priorities between activities and goals (Inzlicht et al., 2014).

Related to the diversity of self-control measures, the exact role of an inhibitory mechanism (vs. selective attention/attentional control) has been questioned in regard to many behavioural measures of inhibition-related executive functioning (Egner & Hirsch, 2005). Furthermore, while some analyses have suggested that a common latent factor unites measures of inhibitory executive functioning (Miyake et al., 2000; Miyake & Friedman, 2012), other research indicates a lack of convergence between these measures (Egner, 2008). As each of our studies assesses the link between reported self-control and one executive functioning task, we are not able to assess links between introspective reports of the ability to control impulses and any latent executive

functioning factor that is common to the Stroop and Flanker tasks. Future work should explore this possibility.

Another limitation is the previously mentioned reliability paradox (Hedge et al., 2017): Robust cognitive tasks do not produce reliable individual differences, making their use as trait-level correlational tools problematic. Behavioural tasks only become well established when between-subject variability is low; however, low between-subject variability hurts reliability for individual differences and deflates correlations. One potential solution, albeit controversial, is to disattenuate correlations undermined by low reliabilities (Muchinsky, 1996). When we disattenaute the meta-analytic correlations, which range from $r = .03–06$, they increase somewhat to $r = .05–.08$. Though slightly increased, this disattenuation suggests that the correlations between reported self-control and inhibition-related executive function is not low because of poor reliabilities, but because they are actually uncorrelated, with less than 0.7% in shared variance. Furthermore, it should be noted that while non-difference scores from the Stroop task (e.g., mean reaction time on incompatible trials) demonstrated good reliability, Bayesian correlations supported a null relationship with reported self-control (see online supplemental materials). Together with the disattenuated correlations, these results suggest that the current findings are unlikely a direct result of the poor reliability in the executive functioning tasks.

Conclusion

The current findings are consistent with a null relationship between performance measures of inhibition-related executive functioning (the Stroop and flanker tasks), the Self-Control Scale, and related measures of self-discipline (Grit, Conscientiousness). Our results highlight empirical and conceptual problems with self-control as a psychological construct, where widely used and established measures of self-control are largely unrelated to each other.

Data Accessibility Statement

The data and scripts are available on our OSF page (https://osf.io/8etus/).

Notes

[1] In addition to the classic Stroop effect, we manipulated the proportion of congruent to incongruent trials in the Stroop task (cf., Logan & Zbrodoff, 1979). However, this manipulation was not modelled as it is not central to the present work. The task comprised 576 trials divided equally into 9 blocks. Blocks were divided into groups of three that varied in the ratio of compatible to incompatible trials (75% compatible/25% incompatible; 50% compatible/50% incompatible; 25% compatible; 75% incompatible). Condition order was counterbalanced between participants, while the three blocks with equal proportions were always presented together. Proportion congruency conditions were collapsed for all analyses. It should be noted that neither the self-control scale nor the self-discipline composite measure predicted the Stroop effect in the majority of conditions (all $rs < −.144$). There was a small negative correlation between the Stroop effect and error rates during the majority compatible condition and the self-discipline composite ($r = −.167$, $p = .044$). However, this correlation should be interpreted with some caution given both the high p-value and the considerable number of comparisons undertaken in this analysis.

[2] We replicated our Bayesian analyses using the non-differences scores (i.e., RT and error-rates on compatible and incompatible trials), and these are presented in the online supplementary materials. The results for the non-difference scores were mixed. In reaction times, we tended to see small positive correlations between reported self-control and performance on both compatible and incompatible trials, whereas smaller negative correlations were observed for error rates on the same trial types. This pattern of results is more consistent with reported self-control correlating with a slight speed-accuracy trade-off (slower overall RT and reduced error-rates), rather than an increase in inhibition-related executive functioning.

[3] A sensitivity analysis using other reasonable choices of priors revealed no concerns that affect the conclusions of the present analyses.

[4] One previous factor analytical investigation suggested that the brief Self-Control Scale can be further divided into subscales with of items reflecting initiatory and inhibitory self-control (de Ridder, de Boer, Lugtig, Bakker, & van Hooft, 2011). Bayesian correlations conducted in JASP supported a null relationship between the inhibition related items and the Stroop effect in RT($r = .022$, $BF_{01} = 14.93$) and error-rates ($r = .077$, $BF_{01} = 23.86$).

[5] All results were identical when no exclusions were applied (see supplemental materials, https://osf.io/jws4x/).

[6] Studies 2b and 2c had high rates of exclusion. While we cannot tell exactly why exclusion rates were so high in these studies, we think that the poor data quality likely arose from the use of online survey companies in which participant panels likely had less experience with behavioural tasks than Mturk participants. It is important to note, however, that our overall conclusions were already supported from the results of study 1 and 2a without including studies 2b–c. However, we opted to include the later studies (after removing poorly performing participants) for the sake of transparent reporting.

[7] Bayesian correlations conducted in JASP supported a null relationship between the inhibition related items of the brief Self-Control Scale (cf., de Ridder et al., 2011) and the Stroop effect in RT($r = −.109$, $BF_{01} = 2.041$) and error-rates ($r = −.052$, $BF_{01} = 5.581$).

Science Foundation's Methods, Measurements, and Statistics panel.

Competing Interests
The authors have no competing interests to declare.

Author Contributions
Saunders, Milyavskaya, Randles, and Inzlicht conceived the study. Saunders, Randles, Inzlicht and Milyavskaya oversaw the design and data-collection for each of the studies. Data was structured and aggregated by Saunders and Randles. Data analyses were conducted by Saunders, Milyavskaya, and Etz. Saunders drafted the initial manuscript and all authors provided critical feedback and edits on the manuscript.

References
Allan, J. L., Johnston, M., & **Campbell, N.** (2010). Unintentional eating. What determines goal-incongruent chocolate consumption? *Appetite, 54*(2), 422–425. DOI: https://doi.org/10.1016/j.appet.2010.01.009

Allom, V., Panetta, G., Mullan, B., & **Hagger, M. S.** (2016). Self-report and behavioural approaches to the measurement of self-control: Are we assessing the same construct? *Personality and Individual Differences, 90,* 137–142. DOI: https://doi.org/10.1016/j.paid.2015.10.051

Baumeister, R. F. (2014). Self-regulation, ego depletion, and inhibition. *Neuropsychologia, 65,* 313–319. DOI: https://doi.org/10.1016/j.neuropsychologia.2014.08.012

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

Bogg, T., & **Roberts, B. W.** (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin, 130,* 887. DOI: https://doi.org/10.1037/0033-2909.130.6.887

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & **Cohen, J. D.** (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652. DOI: https://doi.org/10.1037/0033-295X.108.3.624

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Riddell, A., et al. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software, 20,* 1–37.

Crump, M. J., McDonnell, J. V., & **Gureckis, T. M.** (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One, 8,* e57410. DOI: https://doi.org/10.1371/journal.pone.0057410

Davidson, M. C., Amso, D., Anderson, L. C., & **Diamond, A.** (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44,* 2037–2078. DOI: https://doi.org/10.1016/j.neuropsychologia.2006.02.006

de Ridder, D. T., de Boer, B. J., Lugtig, P., Bakker, A. B., & **van Hooft, E. A.** (2011). Not doing bad things is not equivalent to doing the right thing: Distinguishing between inhibitory and initiatory self-control. *Personality and Individual Differences, 50*(7), 1006–1011. DOI: https://doi.org/10.1016/j.paid.2011.01.015

de Ridder, D. T., 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. DOI: https://doi.org/10.1177/1088868311418749

Duckworth, A., & **Gross, J. J.** (2014). Self-control and grit related but separable determinants of success. *Current Directions in Psychological Science, 23,* 319–325. DOI: https://doi.org/10.1177/0963721414541462

Duckworth, A. L., & **Kern, M. L.** (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality, 45,* 259–268. DOI: https://doi.org/10.1016/j.jrp.2011.02.004

Duckworth, A. L., Peterson, C., Matthews, M. D., & **Kelly, D. R.** (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92* DOI: https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A., & **Seligman, M. E.** (2006). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16,* 939–944. DOI: https://doi.org/10.1111/j.1467-9280.2005.01641.x

Duckworth, A. L., & **Quinn, P. D.** (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment, 91,* 166–174. DOI: https://doi.org/10.1080/00223890802634290

Edmonds, G. W., Bogg, T., & **Roberts, B. W.** (2009). Are personality and behavioral measures of impulse control convergent or distinct predictors of health behaviors? *Journal of Research in Personality, 43,* 806–814. DOI: https://doi.org/10.1016/j.jrp.2009.06.006

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences, 12,* 374–380. DOI: https://doi.org/10.1016/j.tics.2008.07.001

Egner, T., & **Hirsch, J.** (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience, 8,* 1784–1790. DOI: https://doi.org/10.1038/nn1594

Eriksen, B. A., & **Eriksen, C. W.** (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Attention, Perception, & Psychophysics, 16,* 143–149. DOI: https://doi.org/10.3758/BF03203267

Etz, A., & **Vandekerckhove, J.** (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS One, 11*(2), e0149794. DOI: https://doi.org/10.1371/journal.pone.0149794

Etz, A., & **Vandekerckhove, J.** (In press). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin & Review.* DOI: https://doi.org/10.3758/s13423-017-1262-3

Fleming, K. A., Heintzelman, S. J., & Bartholow, B. D. (2016). Specifying associations between conscientiousness and executive functioning: Mental set shifting, not prepotent response inhibition or working memory updating. *Journal of Personality*, *84*, 348–360. DOI: https://doi.org/10.1111/jopy.12163

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*, e109019. DOI: https://doi.org/10.1371/journal.pone.0109019

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., Schmeichel, B. J., et al. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology*, *92*, 325–336. DOI: https://doi.org/10.1037/0022-3514.92.2.325

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*, 1216. DOI: https://doi.org/10.1037/0022-3514.59.6.1216

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504–528. DOI: https://doi.org/10.1016/S0092-6566(03)00046-1

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Steingroever, H., et al. (2017). A tutorial on bridge sampling. *arXiv preprint* arXiv:1703.05984.

Gronau, Q. F., Singmann, H., & Wagenmakers, E. -J. (2017). *Bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors*. Retrieved from: https://github.com/quentinngronau/bridgesampling (R package version 0.2-2).

Hall, P. A. (2012). Executive control resources and frequency of fatty food consumption: Findings from an age-stratified community sample. *Health Psychology*, *31*, 235. DOI: https://doi.org/10.1037/a0025407

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 1–21.

Hofmann, W., Schmeichel, B. J., & Baddeley, A. D. (2012). Executive functions and self-regulation. *Trends in Cognitive Sciences*, *16*, 174–180. DOI: https://doi.org/10.1016/j.tics.2012.01.006

Inzlicht, M., & Gutsell, J. N. (2007). Running on empty: Neural signals for self-control failure. *Psychological Science, 18*, 933–937. DOI: https://doi.org/10.1111/j.1467-9280.2007.02004.x

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, *18*, 127–133. DOI: https://doi.org/10.1016/j.tics.2013.12.009

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality, 44*, 501–511. DOI: https://doi.org/10.1016/j.jrp.2010.06.005

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research, 2*, 102–138.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78–89. DOI: https://doi.org/10.1016/j.jrp.2014.05.003

Kern, M. L., & Friedman, H. S. (2008). Do conscientious individuals live longer? A quantitative review. *Health Psychology*, *27*, 505. DOI: https://doi.org/10.1037/0278-6133.27.5.505

Kopp, B., Rist, F., & Mattler, U. W. E. (1996). N200 in the flanker task as a neurobehavioral tool for investigating executive control. *Psychophysiology*, *33*, 282–294. DOI: https://doi.org/10.1111/j.1469-8986.1996.tb00425.x

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition, 7*, 166–174. DOI: https://doi.org/10.3758/BF03197535

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E. -J. (2017). Replication Bayes factors from Evidence Updating. *PsyArXiv preprints.* Retrieved from: https://osf.io/preprints/psyarxiv/u8m2s/. DOI: https://doi.org/10.31234/osf.io/u8m2s

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science, 244*, 933–938. DOI: https://doi.org/10.1126/science.2658056

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. DOI: https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100. DOI: https://doi.org/10.1006/cogp.1999.0734

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Sears, M. R., 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. DOI: https://doi.org/10.1073/pnas.1010076108

Molden, D. C., Hui, C. M., Scholer, A. A., Meier, B. P., Noreen, E. E., D'Agostino, P. R., & Martin, V. (2012). Motivational versus metabolic effects of carbohydrates on self-control. *Psychological Science, 23*, 1137–1144. DOI: https://doi.org/10.1177/0956797612439069

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*, 63–75. DOI: https://doi.org/10.1177/0013164496056001004

Nigg, J. T. (2001). Is ADHD a disinhibitory disorder? *Psychological bulletin*, *12*, 571. DOI: https://doi.org/10.1037/0033-2909.127.5.571

Pronk, T. M., Karremans, J. C., & **Wigboldus, D. H.** (2011). How can you resist? Executive control helps romantically involved individuals to stay faithful. *Journal of personality and social psychology*, *100*, 827. DOI: https://doi.org/10.1037/a0021993

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & **Meints, J.** (2009). Conscientiousness. In: Leary, M. R., & Hoyle, R. H. (Eds.), *Handbook of individual differences in social behaviour*, 369–381. New York, NY, US: Guilford Press.

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & **Hill, P. L.** (2014). What is conscientiousness and how can it be assessed? *Developmental Psychology*, *50*, 1315. DOI: https://doi.org/10.1037/a0031109

Stan Development Team. (2017). Stan: A C++ library for probability and sampling. Online: http://mc-stan.org.

Streeter, C. C., Terhune, D. B., Whitfield, T. H., Gruber, S., Sarid-Segal, O., Silveri, M. M., Renshaw, P. F., et al. (2008). Performance on the Stroop predicts treatment compliance in cocaine-dependent individuals. *Neuropsychopharmacology*, *33*, 827. DOI: https://doi.org/10.1038/sj.npp.1301465

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. DOI: https://doi.org/10.1111/j.0022-3506.2004.00263.x

Verleger, R., Kuniecki, M., Möller, F., Fritzmannova, M., & **Siebner, H. R.** (2009). On how the motor cortices resolve an inter-hemispheric response conflict: An event-related EEG potential-guided TMS study of the flankers task. *European Journal of Neuroscience*, *30*, 318–326. DOI: https://doi.org/10.1111/j.1460-9568.2009.06817.x

Von Culin, K. R., Tsukayama, E., & **Duckworth, A. L.** (2014). Unpacking grit: Motivational correlates of perseverance and passion for long-term goals. *The Journal of Positive Psychology*, *9*, 306–312. DOI: https://doi.org/10.1080/17439760.2014.898320

Wagenmakers, E. J., Verhagen, J., & **Ly, A.** (2015). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*, 413–426. DOI: https://doi.org/10.3758/s13428-015-0593-0

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H. J., & **Ettinger, U.** (2013). Reliability and plasticity of response inhibition and interference control. *Brain and Cognition*, *81*, 82–94. DOI: https://doi.org/10.1016/j.bandc.2012.09.010