Journal of Research in Personality 88 (2020) 104000

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Do behavioral measures of self-control assess construct-level variance?

Parker A. Dreves *, Ginette C. Blackhart, Matthew T. McBee

East Tennessee State University, United States

ARTICLE INFO

Article history:
Received 15 March 2020
Revised 12 July 2020
Accepted 13 July 2020
Available online 18 July 2020

Keywords:
Self-control
Impulse inhibition
Delay of gratification
Willpower
Measurement

ABSTRACT

A wide range of measures have been used to assess self-control including executive function tasks, delay of gratification tasks, and persistence and willpower tasks. The current study sought to examine the convergent and predictive validity of these measures, provide theoretical backing for why we might or might not expect high correlations between different indicators of the construct, and question whether such measures are assessing construct-level variance. The results largely replicated prior research, with the majority of correlations being small in magnitude and non-significant. Possible interpretations include indicators assessing distinct and unrelated subdomains of self-control, the inappropriate use of measures that maximize within person variance, indicators being plagued by large sources of error variance, or some combination of these.

The benefits of high self-control have been discussed extensively in the psychological literature and range from better emotional stability, school and work performance, and social competence to decreased risk of drug use, obesity, gambling, risky sexual behaviors, and other risk taking behaviors (Duckworth, Tsukayama, & Kirby, 2013; Mischel & Ebbesen, 1970; Reynolds, Richards, Horn, & Karraker, 2004; Sharma, Markon, & Clark, 2014; Tangney, Baumeister, & Boone, 2004). Despite the purported value of this construct, there remains lack of clarity regarding (a) how to best assess the construct in laboratory settings and (b) the ontological nature of the construct. The main goal of this paper is to evaluate commonly used behavioral measures of self-control in terms of indicator intercorrelations, indicator correlations with self-report measures, and indicator correlations with theoretical outcomes of the construct. In doing so, we hope to provide insight on whether these measures are assessing construct-level variance.

Fundamentally, self-control is defined by the prioritization of long-term goals over the near-term temptations that conflict with those goals (Fujita, 2011). Successful self-control is defined as engaging in behaviors that progress one toward a long-term goal or, conversely, abstaining from behaviors that thwart the realization of long-term goals. Although many consider the second part of this definition – abstaining – to be the defining feature of self-control, it should be noted that impulse inhibition is but one strategy among many for furthering long-term goals (Magen & Gross, 2010). Prior research suggests that there are two main types of self-control: inhibitory self-control and initiatory self-control (de Ridder, de Boer, Lugtig, Bakker, & van Hooft, 2011). Inhibitory self-control refers to abstaining from behaviors that conflict with long-term goals, such as resisting the impulse to indulge in unhealthy behaviors that conflict with health goals. By contrast, initiatory self-control refers to enacting some behavior that promotes a long-term goal, such as exercising to promote good health. Factor analysis suggests that inhibitory and initiatory self-control are separable constructs and inhibitory self-control is more predictive of the frequency of undesired behaviors, such as smoking cigarettes or drinking, whereas initiatory self-control is more predictive of desirable behaviors, such as hours of study (de Ridder et al., 2011). This highlights the fact that self-control is a multidimensional construct not just limited to the inhibition of impulses.

Self-control can be subdivided even further into specific strategies for goal-pursuit including situation selection, situation modification, cognitive change, distraction, impulse inhibition, or forming implementation intentions (Baumeister, 2002; Diamond, 2013; Fujita, 2011; Gollwitzer, 1999; Magen & Gross, 2010). Given the variety of ways that self-control may be expressed, it is perhaps unsurprising that self-control has been assessed using seemingly disparate measures including self-report measures (Patton, Stanford, & Barratt, 1995; Tangney et al., 2004), attention and response inhibition tasks (Eriksen & Eriksen, 1974; Nosek & Banaji, 2001; Stroop, 1935; Verbruggen & Logan, 2008), delay of gratification and delay discounting tasks (Frye, Galizio, Friedel, DeHart, & Odum, 2016; Koffarnus & Bickel, 2014; Mischel & Ebbesen, 1970; Richards, Zhang, Mitchell, & de Wit, 1999), and persistence and willpower tasks (Baumeister, Bratslavsky, Muraven, & Tice, 1998). As a result of the diversity among measurement techniques, experts in the field still hotly debate which types of tasks

* Corresponding author.
E-mail address: parkeradreves@gmail.com (P.A. Dreves).

https://doi.org/10.1016/j.jrp.2020.104000
0092-6566/© 2020 Elsevier Inc. All rights reserved.

do and do not utilize self-control resources (Baumeister & Vohs, 2016; Blázquez, Botella, & Suero, 2017; Carter, Kofler, Forster, & McCullough, 2015; Lurquin & Miyake, 2017; Monterosso & Luo, 2013). One explanation for the lack of comparable results among labs using different measures is that each of these measures are assessing different subcomponents of self-control, yet all being referred to imprecisely as "self-control".

The crux of the issue is that grouping all of these strategies under the umbrella term "self-control" makes it unclear as to which processes are actually at work in a given test. Tasks like the Stroop task, the go/no-go task, delay of gratification tests, the cold pressor task, food taste tests, and persistence tasks may allow or disallow the use of different self-control strategies. For example, distraction reliably increases delay times on delay of gratification tasks (Mischel, Ebbesen, & Raskoff Zeiss, 1972; Peake, Hebl, & Mischel, 2002; Sethi, Mischel, Aber, Shoda, & Rodriguez, 2000), but distraction is not a viable strategy when completing measures of executive function such as the Stroop task. Although the different strategies may be important subcomponents of self-control, considering any one of them to be an indicative test of global self-control may be an error. Table 1 summarizes some common behavioral measures of self-control and which subdomain of the construct each may assess.

Importantly, the theoretical model under which one is operating largely determines the pattern of expected relationships among indicators of self-control. A notable distinction here is the difference between reflective constructs and formative constructs. Reflective constructs are constructs that are proposed to exist independently of measurement and can be estimated though their effect on a variety of indicators. In other words, causality flows from the construct (self-control) to the indicators. By contrast, a formative construct is a construct that is defined by a particular set of indicators but itself has no causal influence over its indicators. In a formative model the latent variable is an index number, not an estimate of an underlying factor, and resembles something more like socioeconomic status, the Dow Jones Industrial Average, or the Air Quality Index (van der Maas, Kan, & Borsboom, 2014).

The reflective/formative distinction is important for calibrating our expectations in terms of the convergent and predictive validity of these measures. Supposing a reflective model, substantial correlations would be expected among indicators due to indicators sharing a common cause (Coltman, Devinney, Midgley, & Venaik, 2008). This would be the case for something like intelligence, where *g* is thought to explain variance in abilities like processing speed, retrieval ability, and visual perception (Carroll, 1993). By contrast, in a formative model where indicators are not hypothesized to share a common cause, there is no a priori expected pattern of correlations between indicators. Relevant to this point, research on impulsivity suggests four separable subdomains including premeditation, urgency, sensation seeking, and perseverance -- some of which are completely uncorrelated (as is the case for sensation seeking and premeditation; Whiteside & Lynam, 2001). Regarding the predictive validity of construct indicators, under a reflective model one would expect indicators to predict the same outcomes of the construct. In more concrete terms, if self-control theoretically leads to a particular outcome and several indicators measure self-control, then each indicator should also predict that outcome. By contrast, if self-control is a formative construct, this would not need to be the case. It could be the case that certain indicators only predict certain outcomes and not others.

Keeping these points in mind, we briefly review prior research on the convergent and predictive validity of behavioral measures of self-control. Duckworth and Kern (2011) conducted a meta-analysis looking at relationships between self-report measures, impulse inhibition measures, and delay of gratification tasks. Across the 282 studies included, the authors report a medium effect size ($r_{\text{random}} = 0.27$ [95% CI = 0.24, 0.30]; $r_{\text{fixed}} = 0.34$ [0.33, 0.35]). However, it should be noted that this includes correlations between measures of the same type (e.g. self-report correlated with self-report). Looking at only the correlations between different types of indicators, the correlations were much lower (executive function tasks and delay tasks, $r = 0.11$; executive function and self-report, $r = 0.10$; delay of gratification and self-report, $r = 0.15$). With the exception of self-report measures, even among tasks of the same type correlations were relatively low (executive function and executive function, $r = 0.15$; delay of gratification and delay of gratification, $r = 0.21$). A similar meta-analysis by Cyders and Coskunpinar (2011) also found relatively unimpressive correlations between various measures of impulsivity ranging from $r = 0.09$ to $r = 0.13$ across the 27 studies included. Of note, this study concluded that although there was a small amount of shared variance, the mismatch between self-report and behavioral measures could be suggestive of them assessing different underlying constructs.

Regarding the predictive validity of these indicators, Sharma et al. (2014) conducted a meta-analysis examining relationships between self-report measures of impulsivity, laboratory tests of impulsivity, and daily life behaviors. Behavioral measures included were the Stroop task, the go/no go task, the stop signal task, the Iowa gambling task, and the Wisconsin card sorting task. The daily life outcomes assessed included measures of drug, alcohol and tobacco use as well as aggression, gambling, and risky sexual behaviors. The meta-analysis revealed that, on the whole, behavioral measures were weakly or non-significantly related to daily life behaviors. One exception was the Stroop task, which correlated with alcohol use, aggression, and gambling. Additionally, they found that behavioral measures of impulsivity loaded on different factors they labelled inattention, inhibitory dyscontrol, impulsive decision making, and set shifting. The lack of predictive validity of these indicators is further supported by a large number of independent studies with similar findings (Crean, de Wit, & Richards, 2000; Fine, Steinberg, Frick, & Cauffman, 2016; Lane, Cherek, Rhodes, Pietras, & Tcheremissine, 2003; White et al., 1994). The one exception here may be a relatively large body of research showing that delay of gratification tasks do seem to predict life outcomes theoretically associated with self-control including smoking, drug use, achievement test scores, GPA, BMI, and emotional stability (Ayduk et al., 2000; Duckworth et al., 2013; Mischel & Ebbesen, 1970; Mischel et al., 1972; Mischel, Shoda, & Peake, 1988; Shoda, Mischel, & Peake, 1990; Mischel, Shoda, &

Table 1
A summary of measures of state self-control.

Measures	Examples	Factors assessed
Executive function tasks	Stroop, go/no go, stop-signal, flanker	Cognitive inhibition, resistance to distractor interference, attention, reaction time
Delay of gratification/ delay discounting tasks	Monetary or food related delay of gratification, adjusting amount	Preference for delayed rewards, rate of delay discounting (*k*)
Persistence/willpower tasks	Cold Pressor, taste test, handgrip, impossible puzzle, bitter drink	Behavioral inhibition, behavioral initiation, pain tolerance

Rodriguez, 1989; Reyna & Wilhelms, 2017; Reynolds, Ortengren, Richards, & de Wit, 2006; Reynolds et al., 2004; Watson & Milfont, 2017). With these considerations in mind, the primary goal of the present study is to examine the convergent and predictive validity of commonly used measures of self-control and, in doing so, provide insight into whether these measures are assessing construct-level variance.

1. Method

1.1. Power analysis

Based on research by Duckworth and Kern (2010), Saunders, Milyavskaya, Etz, Randles, and Inzlicht (2017), Reynolds et al. (2006), Lane et al. (2003), Schmeichel and Zell (2007) and White et al. (1994), correlations between different indicators ranged from small and non-significant up to as high as $r = 0.35$. Based on this level of inconsistency, we calculated the sample size needed to detect a significant correlation of $r = 0.20$ between any two given indicators. The necessary sample size needed to detect correlations of 0.20 at 80% power is $n = 192$. Power analysis was conducted using the R package "pwr" (Champely, 2018).

1.2. Participants

Participants were 197 undergraduate students recruited from a public university in the southeastern United States. However, data from 6 participants were discarded due to errors in data collection or non-compliance with study instructions, leaving a final sample of 191. The majority of participants were female (61.3% female, 32.5% male, 0.5% transgender, and 5.7% declined to identify), Caucasian (80.1% Caucasian, 6.8% Black, 4.7% Hispanic/Latino, 8.4% other) and had a mean age of 19.7 (SD = 3.7). Research participation was incentivized by offering 3 research credits for participation, which could be exchanged for extra credit in a psychology course.

1.3. Procedure

Participants first completed behavioral measures of self-control in a randomized order. These included the Stroop task, the go/no go task, the flanker task, the cueing task, the handgrip task, the cold pressor, a taste test, the Iowa Gambling Task, and two impossible puzzle tasks. Next, participants completed the survey portion of the study, which included the Self-Control Scale (Tangney et al., 2004; α = 0.800), the Barratt Impulsiveness Scale (Patton et al., 1995; α = 0.841), the AUDIT (Saunders, Aasland, Babor, de la Fuente, & Grant, 1993; α = 0.787), an impulse spending questionnaire (Güre, 2012; α = 0.887), the Health-Related Behavior Questionnaire (Balding, 2008; α = 0.750 for healthy food; α = 0.621 for unhealthy food), and the Leisure Time Exercise Questionnaire (Godin & Shephard, 1985; α = 0.599). GPA was recorded via having participants log into their university account. See Table 2 and Table 3 for descriptive statistics and reliability coefficients for all self-report and behavioral measures. Finally, delay of gratification was assessed at the end of the study. This had to be assessed last due to the nature of the focal reward being used, as explained below.

1.4. Executive function

Four executive function tasks were used, all of which were programmed and administered using PsyToolkit (Stoet, 2017), a digital platform for developing psychological tests. The four executive function tasks were the Stroop task (Stroop, 1935), the go/no go task (Nosek & Banaji, 2001), the flanker task (Eriksen & Eriksen, 1974) and a spatial cueing task (Posner, 1980). We administered 150 trials of the Stroop task (75 congruent, 75 incongruent), 100 trials of the go/ no go task (30 no go/ 70 go), 150 Flanker trials (75 congruent, 75 incongruent), and 100 trials of the cueing task (30 invalid cues). Importantly, for the go/ no go tasks, the outcome measure was the count of trials on which the participant was not supposed to respond but did respond (e.g., errors of commission). This decision was made based on previous research showing that errors of omission relate to inattention, whereas errors of commission relate significantly to symptom counts of impulsivity (Bezdjian, Baker, Lozano, & Raine, 2009).

1.5. Delay of gratification

The Monetary Choice Questionnaire (Kirby, Petry, & Bickel, 1999) was used to assess individual's willingness to delay immediate gratification in favor of larger rewards. This is a 27-item questionnaire wherein participants must choose between hypothetical rewards delayed over varying amounts of time. k, or an individual's rate of delay discounting, was calculated using R code developed by Gray, Amlung, Palmer, and MacKillop (2016). k ranges from 0.00016 to 0.25 with lower values representing lower rates of delay discounting, or higher self-control. This scale shows excellent reliability in the current study (α = 0.928).

In addition to the delay discounting task, participants also completed a delay of gratification task designed to resemble the experiments carried out by Mischel (1958,1961) (1970; 1972; 1988; 1989) wherein participants were presented with an option between a less desirable but immediate reward or a more desirable but delayed reward. Because the sample in the current study was comprised of college students, the focal reward used was not food but instead research credits. The decision to not use food was made because adults vary widely in the intensity of their food and weight related goals, meaning food would not elicit a goal conflict of equal strength among participants. By contrast, it is a comparatively safer assumption that participants enrolled in our study were motivated to obtain research credits as this constitutes the primary way that research participation is incentivized.

To this end, participants were presented with three options at the end of the study. The first option (immediate gratification option) was to be awarded the three research credits they had earned and be allowed to leave immediately. The second option

Table 2
Descriptive statistics for self-report measures.

Measure	Mean	SD	Min	Max	Reliability
Self-Control Scale	3.26	0.60	1.69	4.77	α = 0.800
Barratt Impulsiveness	2.91	0.37	1.53	3.87	α = 0.841
Impulse Shopping	1.30	0.38	1.00	5.00	α = 0.887
AUDIT	1.30	0.83	1.00	2.70	α = 0.787
Exercise	2.49	1.04	1.00	5.00	α = 0.599
Healthy Food	4.69	1.04	1.60	7.00	α = 0.750
Unhealthy Food	3.98	1.04	1.50	6.50	α = 0.621

Table 3
Descriptive statistics for behavioral measures.

Measure	Mean	SD	Min	Max
Stroop	113.73 (ms)	64.81 (ms)	−107.00 (ms)	298.00 (ms)
Flanker	25.97 (ms)	39.37 (ms)	−112.00 (ms)	130.00 (ms)
Cueing	71.23 (ms)	55.13 (ms)	−287.00 (ms)	291.00 (ms)
Go/no go	3.71	2.94	0	15
Monetary Choice (k)	0.02319	0.03605	0.00015	0.24942
Handgrip	17.98 (s)	15.77 (s)	3 (s)	87 (s)
Cold Pressor	97.94 (s)	66.10 (s)	3 (s)	180 (s)
Vegetables Consumed	9.74 (g)	10.75 (g)	0 (g)	50 (g)
Candies Consumed	11.04 (g)	10.33 (g)	0 (g)	50 (g)
Iowa Gambling Task	31.35	17.50	0	87
Anagram Persistence	18.72 (m)	8.40 (m)	2 (m)	30 (m)
Math Persistence	12.23 (m)	6.98 (m)	1.6 (m)	30 (m)

(intermediate gratification) was to stay in the study for an additional half-hour completing extra surveys but receive an additional half research credit (3.5 total) in return. The third option (delayed gratification) was to stay in the study for an additional hour of surveys but receive an additional full research credit (4 total) in return. This directly assessed participants' willingness to sacrifice an immediate desire (leaving early) in favor of a longer-term goal (extra credit in a course). To ensure that participants' decisions were not affected by scheduling conflicts, all participants were told to allow at least three hours for the study (the protocol without the possible additional 1 h of tasks took <2 h). In the current sample, 54.5% chose to leave immediately, 11% chose to stay for an extra half credit, and 34.5% chose to stay for the additional full credit. This exact methodology has been used in previous research and has been found to be sensitive to changes in goal strength (Dreves & Blackhart, 2019).

1.6. Probability discounting

Probability discounting was assessed via the Iowa Gambling Task, a task that was first developed to assess impulsivity in people with damage to the prefrontal cortex (Bechara, Damasio, Damasio, & Anderson, 1994) but has subsequently been used to assess self-control (Sharma et al., 2014). This task assesses impulsive behavior through people's decisions to draw from a high-risk or low risk deck of cards. Participants completed 100 trials and the number of times they picked from the high-risk deck was counted. Higher scores indicate higher impulsiveness. In the present study, the average number of times participants picked from the high-risk deck was 31.35 (SD = 17.5, min = 0, max = 87).

1.7. Persistence/willpower

We administered five willpower tasks. These were the taste test, the cold pressor, two impossible puzzles, and the handgrip. The taste test was a measure of calorie consumption and has been used as a dependent measure of self-control in many studies of ego-depletion (Hagger, Wood, Stiff, & Chatzisarantis, 2010). This was performed identically to how it was described in Hagger et al. (2013), since this is representative of the way this has been carried out in most ego-depletion paradigms. For this task, participants were presented with a cover story about market researchers being interested in college students' perceptions of various foods. Participants were presented with two different types of candies (Skittles™ and M&Ms™) as well as two different types of vegetables (raw broccoli and raw cauliflower) and asked to rank them on a variety of dimensions such as taste, texture, and appearance. 50 g of each food were weighed out and presented to the participant. When the participant signaled that they were done evaluating the foods, the experimenter weighed the remaining foods and recorded the amount consumed in grams. In this sample, the average amount of candies consumed was 11.04 g (SD = 10.3, min = 0, max = 50) and the average amount of vegetables consumed was 9.74 g (SD = 10.75, min = 0, max = 50).

For the cold pressor task, participants were told that we would be testing their pain tolerance and were instructed to hold their hand in a bowl of ice water for as long as they could. The water was maintained between 35 degrees and 40 degrees Fahrenheit as advised by Mitchell, MacDonald, and Brodie (2004). The cold pressor is frequently used in psychological studies to simulate pain (Peckerman et al., 1998) and therefore measures a participant's willingness to tolerate an unpleasant stimulus. In theory, this task requires impulse inhibition because individuals must inhibit the impulse to remove their hand from the ice water. To ensure participant safety, anyone who passed the 3-minute mark was instructed to remove their hand from the water. The time (in seconds) was recorded by the experimenter. The average amount of time that participants persisted on the cold pressor was 97.94 s (SD = 66.1, min = 3, max = 180).

Participants also completed the handgrip task, which has been used in ego-depletion paradigms (Hagger et al., 2010; Muraven, Tice, & Baumeister, 1998). For this task, participants were given a handgrip and told squeeze it for as long as possible. To account for individual differences in maximum grip strength, participants first completed a baseline measure of grip strength using a dynamometer. The handgrip was then calibrated to match their maximum grip strength. Thus, this task should have been equally difficult for everyone regardless of their maximum grip strength. Like prior research, a coin was placed between the two grips. When the participant loosened their grip enough that the coin fell out, the experimenter stopped the timer. The average amount of time that participants persisted on the handgrip was 17.98 s (SD = 15.77, min = 3, max = 87).

Finally, persistence was measured with two impossible puzzle tasks. One of them was an extremely difficult mathematics puzzle and the other was a list of unsolvable anagrams. For both tasks, participants were allowed a maximum of 30 min but were told that they could notify the experimenter at any time if they would like to move on. The dependent variable was the amount of time participants persisted before giving up. The average amount of time participants spent on the anagrams was 18.72 min (SD = 8.4, min = 2, max = 30), and the average amount of time spent on the math puzzle was 12.22 min (SD = 6.98, min = 1.3, max = 30).

1.8. Data cleaning and normalization

Of note, there were several extreme scores present in the data. For example, although 90% of participants scored between 4.6 and 52.8 on the handgrip, a small number of participants exceeded 80 s. Therefore, in order to normalize the data, all scores more than

3 interquartile ranges away from the first and third quartiles were capped at 3 IQR above the median. This resulted in five handgrip scores above 64 s being capped at 64, one go no go score above 14 being capped at 14, five vegetable consumption scores over 33 g being capped at 33 g, four candy consumption scores above 48 being capped at 48, and one cueing score above 282 being capped at 282. Finally, to ensure that all variables were scaled similarly and to avoid large discrepancies in covariances, means were converted to Z-scores.

2. Results

2.1. Indicator intercorrelations

Table 4 displays intercorrelations between behavioral indicators of self-control. We used the Benjamini-Hochberg correction for multiple comparisons (Benjamini & Hochberg, 1995), which ranks correlations based on p-values and is less penalizing to lower p-values, thus increasing statistical power. An excel spreadsheet with functions for calculating corrected p-values was retrieved online (McDonald, 2014). Notably, only the six strongest correlations remained significant after controlling for multiple comparisons. Handgrip persistence correlated positively with cold pressor persistence ($r = 0.22$, $p_{\text{adjusted}} = 0.034$), anagram persistence ($r = 0.25$, $p_{\text{adjusted}} = 0.007$), and choices on the delay of gratification task ($r = 0.31$, $p_{\text{adjusted}} = 0.001$). Persistence on the cold pressor correlated positively with the amount of vegetables consumed on the taste test ($r = 0.25$, $p_{\text{adjusted}} = 0.009$). Lastly, anagram persistence correlated positively with math puzzle persistence ($r = 0.38$, $p_{\text{adjusted}} < 0.000$), and vegetables consumed correlated positively with candies consumed ($r = 0.27$, $p_{\text{adjusted}} = 0.003$). On the whole, correlations between indicators ranged from $r = -0.17$ to $r = 0.38$ with an average absolute value of $r = 0.09$. We averaged the absolute value because, theoretically speaking, some correlations were expected to be in the negative direction (e.g., puzzle persistence and risky draws on the gambling task) and others in the positive direction (e.g., puzzle persistence and handgrip persistence). Correlations were converted to Fisher's Z prior to averaging, in line with prior recommendations (Corey, Dunlap, & Burke, 1998).

2.2. Indicator relationships with construct consequences

Table 5 displays correlations between the various indicators and the life outcomes theoretically associated with self-control, and Table 6 displays correlations between self-report measures of self-control and the selected life outcomes. After controlling for multiple comparisons, only cold pressor persistence correlated significantly with exercise ($r = 0.25$, $p_{\text{adjusted}} = 0.038$). Taken together, indicator relationships with construct consequences are inconsistent and weak, at best. Overall, correlations between indicators and construct consequences ranged from $r = -0.18$ to $r = 0.26$ with an average absolute value of $r = 0.07$.

2.3. Indicator relationships with self-report measures

Correlations between the behavioral measures and self-report measures are presented in Table 7. By and large, there were very few statistically significant relationships between the behavioral measures and self-reported self-control. In fact, after applying the Benjamini-Hochberg correction for multiple comparisons, there were no significant relationships between self-report measures and behavioral measures. Correlations between behavioral indicators and self-report measures ranging from $r = -0.18$ to $r = 0.20$ with an average absolute value of $r = 0.08$.

2.4. Factor analysis of behavioral measures

As a final test, we wanted to factor analyze the behavioral measures of self-control. Prior to factor analysis, we applied the Kaiser-Meyer-Olkin (KMO) Test to assess shared variance among indicators. The KMO Test provides a measure of how suited the data are for factor analysis, and low value on this test (<0.60) means that there is a low proportion of shared variance among variables making the data unsuitable for dimension reduction techniques. A high value (closer to 1) means that there is a large proportion of shared variance that may be due to underlying factors (Kaiser, 1974). Using all thirteen indicators of self-control yielded a KMO value of 0.52. Kaiser (1974) classified values falling between 0.50 and 0.59 as "miserable", suggesting that these data are not suited well for factor analysis. Indeed, the scree plot (Fig. 1) does not show any sign of a clear levelling off point, suggesting that there is not a shared factor(s) that explains variance across these items. See Table 8 for factor loadings.

3. Discussion

3.1. General discussion

The goal of the current research was to add to the body of literature regarding the convergent and predictive validity of behavioral measures of self-control and, in doing so, provide insight on whether these measures are assessing variance at the construct level. By and large, we replicated previous research by finding that that indicator intercorrelations among behavioral measures are

Table 4
Correlations between the behavioral measures of self-control.

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Stroop	–												
2. Flanker	0.04	–											
3. Cueing	0.01	−0.08	–										
4. Go/no go	0.06	-0.15^{*}	0.09	–									
5. Iowa	0.20^{**}	−0.05	0.17^{*}	0.07	–								
6. Math Puzzle	−0.01	0.09	-0.17^{*}	0.03	−0.09	–							
7. Anagram	−0.11	0.05	−0.06	0.02	−0.02	**0.38****	–						
8. Cold Pressor	−0.05	−0.04	−0.02	−0.05	−0.03	0.03	0.14	–					
9. Handgrip	-0.15^{*}	−0.07	−0.01	0.12	−0.09	0.10	**0.25****	**0.22****	–				
10. Candies	−0.06	0.05	−0.04	−0.07	0.17^{*}	0.07	−0.03	0.08	−0.06	–			
11. Vegetables	0.09	0.07	−0.05	−0.03	−0.05	0.20^{**}	0.12	**0.25****	0.09	**0.27****	–		
12. Discount *k*	0.00	−0.05	−0.06	0.04	−0.02	0.04	−0.11	−0.07	−0.08	−0.09	−0.06	–	
13. Delay Task	−0.09	0.02	−0.01	−0.11	0.01	−0.03	0.19^{**}	−0.02	**0.31****	−0.07	0.01	−0.07	–

$^{*}p < .05$, $^{**}p < .01$; Note: After applying the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) for multiple comparisons (setting the FDR to 0.05 and entering the 78 correlations between the 13 behavioral measures), 6 correlations remained significant (bolded).

Table 5
Correlations between behavioral indicators and life outcomes.

Measure/life outcome	GPA	AUDIT	Impulse shopping	Healthy foods	Unhealthy foods	Exercise
Stroop	0.11	−0.08	−0.03	0.12	−0.03	−0.09
Flanker	−0.03	−0.12	0.03	0.05	0.06	−0.06
Cueing	0.00	−0.12	0.00	0.05	−0.02	0.10
Go/no go	0.03	0.01	0.19**	−0.03	0.01	0.07
Iowa	0.07	−0.04	0.20**	−0.01	0.06	0.05
Math Puzzle	0.02	0.05	0.07	−0.06	0.01	−0.05
Anagram	−0.11	0.00	−0.06	0.07	−0.08	−0.02
Cold Pressor	−0.05	0.00	−0.12	0.17*	−0.17*	**0.25** **
Handgrip	−0.02	0.06	−0.15*	0.08	−0.12	0.08
Candies	0.01	0.10	0.05	−0.20*	0.06	0.11
Vegetables	−0.05	0.03	−0.03	0.11	−0.15*	0.11
Discount k	0.16*	0.02	0.07	0.08	−0.08	0.12

*p < .05, ** p < .01; Note: only the correlation between cold pressor persistence and self-reported exercise remained significant after applying the Benjamini-Hochberg correction (setting FDR to 0.05 and entering in the 78 correlations between behavioral measures and life outcomes).

Table 6
Correlations between self-report measures and life outcomes.

Measure	1	2	3	4	5	6	7	8
1. SCS	–							
2. BIS	0.60**	–						
3. GPA	0.11	0.09	–					
4. AUDIT	−**0.23** **	−0.16*	−0.06	–				
5. Impulse Shop	.−**0.32** **	−**0.45** **	0.04	0.01	–			
6. Health Food	0.16*	0.15*	0.02	0.13	−0.08	–		
7. Unhealthy Food	−0.05	−0.16*	0.07	−0.03	0.23**	−0.31**	–	
8. Exercise	0.02	−0.10	0.02	−0.01	−0.06	0.32**	−0.13*	–

*p < .05, **p < .01; Note: After applying the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) for multiple comparisons (setting the FDR to 0.05 and entering the 12 correlations between the 2 self-report measures and the 6 life outcomes), 3 correlations remained significant (bolded).

Table 7
Correlations between behavioral measures and self-report measures.

Measure	SCS	BIS
Stroop	0.01	−0.03
Flanker	0.20**	−0.13
Cueing	0.03	0.05
Go/no go	0.03	−0.03
Iowa	0.02	0.04
Math Puzzle	−0.13	0.16*
Anagram	0.05	−0.09
Cold Pressor	0.06	−0.14
Handgrip	−0.02	−0.11
Candies	−0.17*	0.16*
Vegetables	−0.18*	0.05
Discount k	−0.04	0.11
Delay of Grat.	0.02	−0.10

*p < .05, ** p < .01; Note: no correlations remained significant after applying the Benjamini-Hochberg correction to the 26 correlations in this table (setting the FDR to 0.05).

low (Duckworth & Kern, 2010; Lane et al., 2003; Reynolds et al., 2006; Saunders et al., 2017; Schmeichel & Zell, 2007; White et al., 1994) and that indicators have little relationship with the theoretical consequences of the construct (Fine et al., 2016; Lane et al., 2003; Sharma et al., 2014; White et al., 1994). In particular, we observed a consistent pattern of low indicator intercorrelations ranging from r = −0.17 up to r = 0.38 with an average absolute value of r = 0.09, and only six being significant after controlling for multiple comparisons.

The most optimistic interpretation of these few significant correlations is that these indicators share variance due to an underlying construct. Somewhat notably, all of the significant correlations are in the direction that would be expected by theory (for example, longer delay times correlated positively with handgrip persistence). However, two of these correlations (correlations between the two taste tests and correlations between the two puzzle persistence tasks) are very likely due, at least in part, to method effects (Maul, 2013). Regarding indicator relationships with construct outcomes, correlations ranged from r = −0.18 to r = 0.26 with an average absolute value of r = 0.07 and only one significant correlation after applying the Benjamini-Hochberg correction for multiple comparisons. Finally, indicator relationships with established self-report measures were low, with none being significant after applying the Benjamini-Hochberg correction for multiple comparisons.

We propose four possible explanations for the low correlations found between indicators of self-control both in the present study and numerous others. The first interpretation is that, as discussed previously, self-control could be better conceptualized as a formative construct and thus correlations between indicators should not necessarily be expected. The second interpretation is that many of the behavioral measures used, particularly the EF measures, were designed to assess within-subject effects and may have limited effectiveness in between-subjects designs (Dang, King, & Inzlicht, 2020). The third interpretation is that many of these measures, particularly the ones that have not been thoroughly vetted such as the persistence and willpower tasks, may contain large sources of error variance thus making them unreliable measures and are thus incapable of telling us much about self-control. Finally, the fourth interpretation is that these measures simply do not assess construct-level variance.

Regarding the first interpretation -- that self-control is in fact a formative construct -- we do not think that this explains the low correlations observed in the present study. This is because even is self-control is formative, correlations would still be expected between indicators of the same subdomain (e.g. EF tasks would correlate with other EF tasks, even if EF tasks do not correlate with delay of gratification tasks). The fact that we observed so few significant correlations even among indicators of the same subdomain

Table 8
Factor loadings for behavioral indicators.

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Stroop	−0.14	−0.03	0.08	0.39	0.24	**0.69**
Flanker	−0.06	0.15	−0.07	0.01	**0.75**	−0.05
Cueing	0.18	−0.29	0.02	**0.50**	−0.27	−0.07
Go/no go	−0.29	0.31	−0.18	0.20	**−0.60**	0.09
Iowa	−0.04	0.02	−0.03	**0.84**	0.01	0.01
Math Puzzle	−0.04	**0.81**	0.11	−0.14	0.08	0.06
Anagram	0.40	**0.72**	0.06	0.03	−0.06	−0.11
Cold Pressor	0.20	−0.07	**0.72**	−0.08	−0.15	−0.04
Handgrip	**0.71**	0.11	0.27	−0.09	−0.12	−0.10
Candies	−0.36	0.09	**0.46**	0.27	0.19	**−0.44**
Vegetables	−0.07	0.23	**0.74**	0.04	0.14	0.13
Discount k	−0.07	0.03	0.02	−0.16	−0.20	**0.67**
Delay of Grat.	**0.73**	0.06	−0.13	0.13	0.17	−0.04

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Six factors extracted with Eigenvalues greater than 1. First three factors explain 36% of the variance in scores (Factor 1 = 14.5%, Factor 2 = 11.5%, Factor 3 = 10%).

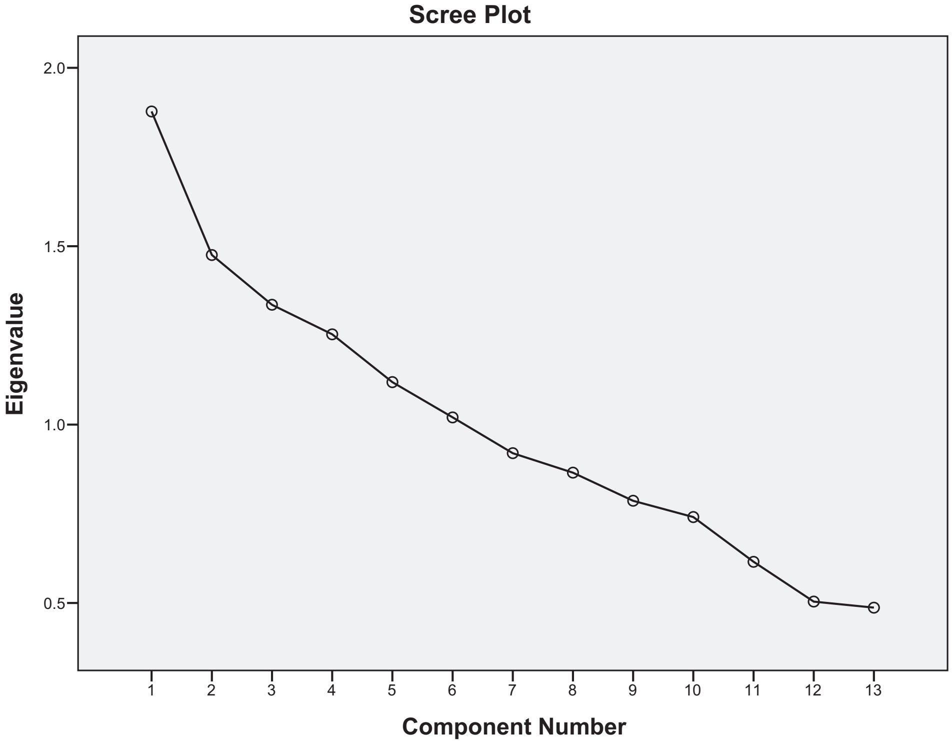

Fig. 1. Scree plot for behavioral indicators.

suggests that the problem lies in the tests themselves, not necessarily the formulation of the construct (of course without proper tests, questions about the formulation of the construct will have to wait). Moreover, after controlling for multiple comparisons, there are no significant relationships between indicators and theoretical outcomes of the construct. This leads us to think that interpretations two, three, or four are more likely candidates for explaining the observed results.

Regarding the second interpretation -- that the behavioral measures used were designed to assess within-subject effects -- a recent article by Dang et al. (2020) highlights the perils associated with using measures that were configured for within-subject designs for between-subjects designs. They argue that tasks like the Stroop task were designed such that everyone shows an interference effect with little variability around the effect, and in fact this is part of what makes the effect so reliable. Given the present results this appears to be the case, as the 95% range for Stroop Scores ranged from 7 ms up to 231 ms. In other words, 95% of the variability in Stroop scores occurred in the span of about 1/5 of a second. It's not all that surprising, then, that such a limited range of scores doesn't tell us all that much about outcomes as varied as substance use, academic performance, and health behavior. A very similar point is made by Hedge, Powell, and Sumner (2018), who noted exceptionally low variance between individuals using measures of this sort, meaning that they are not good tools for running between-subject correlations. Ultimately, we find this a very legitimate criticism of the use of such measures and believe that this does in part explain the lack of relationships, particularly among EF measures.

The third possible explanation for the limited convergent and predictive validity of these measures is simply that they contain large sources of additional variance. Many of these measures may be subject to forces that are difficult or impossible to control for like hunger, participants prior experience, or affinity for certain tasks. For example, it is possible that on the food taste test participants consumption was more a function of how long it had been since they last ate, rather than a function of their self-control. As another example, perhaps athletes who have experience with ice-baths for sports related reasons were better prepared for the cold pressor test (which may in part explain the relationship between self-reported exercise and cold pressor persistence).

Lastly, it could simply be the case that these behavioral measures, by and large, do not assess construct-level variance. We find this interpretation somewhat likely because, thinking back to the definition of self-control, few of these measures actually assess the prioritization of distal motives over proximal motives. Participants have little to gain from holding their hand in ice water or naming the color of a word 20 ms faster. Although pain tolerance and response inhibition may be capacities that are on occasion needed to pursue long-term goals, we believe that they are the means and not the end of self-control. It is possible to imagine an individual with very high impulse inhibition who simply lacks sufficient motivation to use it to resist impulses.

Unfortunately, it is extremely difficult to distinguish this interpretation from interpretation three. Regardless, whichever interpretation is correct – simply not assessing construct level variance or being plagued by error variance – it is probably the case that more valid behavioral measures need to be adopted. Ultimately, we find it likely that the observed low correlations are due to some combination of these latter three explanations. Moreover, if any of these interpretations are correct, then these behavioral measures are not positioned to tell us anything about the structure of the construct itself.

3.2. Limitations

There are a number of limitations of the present research. First, it is possible that many of the laboratory test of self-control used in the present research do not resemble real-life acts of self-control. For example, on the adjusting amount task, hypothetical rewards were used instead of real rewards. This could be relevant because research suggests that participants tend to respond to risk differently on hypothetical tasks as opposed to real tasks, with participants being more risk-sensitive on real tasks (Xu, Fang, & Rao, 2013). Similarly, on tasks like the food taste test, the environment and context of a laboratory does not match that of naturalistic food-related temptations, which more often occur in restaurants or grocery stores. For tasks like the cold pressor, participants may not have had a reason to be motivated to endure the pain of holding their hand in ice water. Usually when people force themselves to endure pain it is in the service of a long-term goal, such as becoming stronger or more athletic, which was not the case in the present study. Relatedly, it should be noted that this research was conducted on undergraduate college students at a university in the southeastern United States, meaning that these measures of self-control may have better or worse effectiveness in different populations (Simons, Shoda, & Lindsay, 2017). For example, these measures may be most effective when used in populations with more severe self-regulation issues, such as among individuals with damage to the prefrontal cortex (Bechara et al., 1994).

Another concern was that performance may be inconsistent across tests due to ego-depletion effects (Baumeister et al., 1998). The ego-depletion effect refers to a phenomenon wherein an initial act of self-control can "deplete" self-control resources, leading to poorer performance on subsequent tasks requiring self-control. Although the effect is somewhat controversial (Carter et al., 2015), this was a concern in the present study due to the quantity of measures being administered. If the ego-depletion effect is a real phenomenon, substantial depletion effects would be expected by the end of this study. To protect against this, the order of the behavioral measures of self-control was randomized. However, due to the nature of the delay of gratification task (offering research credits for extra time in the study), it had to be the last measure administered.

Another limitation of the data is that a few of the tests, in particular the cold pressor task and the anagram task, showed ceiling effects. On the anagram task, 20.9% of participants persisted for the entire thirty minutes. On the cold pressor task, 35.6% of participants persisted for the maximum of 180 s. These ceiling effects reduced the variability in these measures and could have limited the inferences drawn from them. Related to this concern, reliability coefficients are lacking for executive function measures of self-control such as the Stroop, go no go, flanker, and cueing tasks. Unfortunately, the software we used did not output data on every trial but rather automatically calculated and output respondent level means. For example, although participants completed 150 Stroop trials, all that was recorded for each individual participant was the average time on incongruent trials, congruent trials, and their interference effect. Given that the data to compute reliability coefficients for the executive function measures are unavailable to us, this represents a major blind spot of the current research.

3.3. Implications for future research

We find it likely that the available measures either are not assessing variance related to goal-relevant processes, cannot be used for between-subjects designs, and/or contain large sources of error variance. We would strongly suggest that researchers consider new behavioral measure that are more broadly targeted at assessing participants' tendency to prioritize distal goals over proximal goals, rather than the specific strategies they use to do so. This should have the effect of increasing ecological validity, since in real-life acts of self-control participants are not just limited to one self-regulatory strategy but rather have an arsenal of behaviors they can use to meet situation demands. In developing such measures, it will be important to consider how the incentive value of proximal/distal motives varies across participants. The temptation to eat a slice of chocolate cake is not equally strong across all individuals, nor is the desire to lose weight equally strong across all individuals. The impossible puzzle task might have a different meaning for someone who bases their self-concept around their intelligence and ingenuity than someone who bases their self-concept on their appearance or physical ability. The point is that goal conflicts are different for different people and it is unlikely that any one task will equally assess self-control in all individuals. In fact, given the heterogeneity among individuals' objectives and their strategies for reaching those objectives, there may be no "one size fits all" behavioral measure of self-control.

If and when the challenge of developing new behavioral measures of self-control is taken on, it will be of crucial importance to keep in mind what self-control fundamentally is. As defined by Fujita (2011), self-control is "the process of advancing distal rather than proximal motivations when the two compete" (p. 352). Therefore, what any good measure of self-control should assess is an individual's tendency to prioritize distal motives, particularly in situations where proximal motives are competing with these. This aspect balancing short-term rewards and long-term rewards is sorely lacking from current behavioral measures of self-control. For instance, rewards do not increase the longer an individual persists on an impossible puzzle. It is possible that current methods could be modified to induce goal conflicts. For example, imagine offering participants $10 if they were able to solve the impossible puzzle. What one would likely find is that suddenly, a lot more participants would have the "self-control" to persist the entire time. Individuals don't blindly persist on difficult tasks for no reason. Individuals are constantly engaged in cost-benefit analyses, determining whether the level of effort for a given outcome is "worth it".

In conclusion, currently available measures of self-control do not appear to be assessing important aspects of the construct, or at the very least are inappropriate for the between-subjects types of analyses they are so often used for (Dang et al., 2020; Hedge et al., 2018). At best, currently available measures are only assessing some of the means through which self-control may be exercised, but not the construct itself. This is evidenced by the poor convergent, predictive, and concurrent validity found in this study as well as numerous others. We suggest that researchers make attempts to develop new measures that focus less on the specific means which individuals use to exert self-control (e.g., inhibition, distraction, pain tolerance, etc.) and more on the defining feature of the construct itself -- namely the tendency to prioritize distal motives over proximal motives. In doing so, it will also be important to consider how the incentive value of different proximal and distal goals varies across individuals. Until these goals are accomplished, questions regarding the antecedents and consequences of self-control and phenomena such as ego-depletion may be unanswerable. As Cronbach and Meehl (1955) (p. 287), “if the obtained correlation departs from the expectation, however, there is no way to know whether the fault lies in test A, test B, or the formulation of the construct.”

Note

Data cannot be made available due to the IRB data security requirements at the institution.

This research was not preregistered.

References

Ayduk, O., Mendoza-Denton, R., Mischel, W., Downey, G., Peake, P. K., & Rodriguez, M. (2000). Regulating the interpersonal self: Strategic self-regulation for coping with rejection sensitivity. *Journal of Personality and Social Psychology, 79*(5), 776–792. https://doi.org/10.1037/0022-3514.79.5.776.

Balding, J. (2008). *Collecting good data: The primary health related behaviour questionnaire (Version 11.2 Manual)*. Exeter: Schools Health Education Unit.

Baumeister, R. F. (2002). Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of Consumer Research, 28*(4), 670–676. https://doi.org/10.1086/338209.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality And Social Psychology, 74*(5), 1252–1265. https://doi.org/10.1037/0022-3514.74.5.1252.

Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science, 11*(4), 574–575. https://doi.org/10.1177/1745691616652878.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1–3), 7–15. https://doi.org/10.1016/0010-0277(94)90018-3.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1007/978-1-4419-9863-7_1215.

Bezdjian, S., Baker, L. A., Lozano, D. I., & Raine, A. (2009). Assessing inattention and impulsivity in children during the Go/NoGo task. *British Journal Of Developmental Psychology, 27*(2), 365–383. https://doi.org/10.1348/026151008X314919.

Blázquez, D., Botella, J., & Suero, M. (2017). The debate on the ego-depletion effect: Evidence from meta-analysis with the p-uniform method. *Frontiers In Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00197.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511571312.

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*(4), 796–815. https://doi.org/10.1037/xge0000083.

Champely, S. (2018). Pwr: basic functions for power analysis. R package version 1.2-2. https://CRAN.R-project.org/package=pwr

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61*(12), 1250–1262.

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson rs and Fisher’s z transformations. *The Journal of General Psychology, 125*(3), 245–261. https://doi.org/10.1080/00221309809595548.

Crean, J. P., de Wit, H., & Richards, J. B. (2000). Reward discounting as a measure of impulsive behavior in a psychiatric outpatient population. *Experimental And Clinical Psychopharmacology, 8*(2), 155–162. https://doi.org/10.1037/1064-1297.8.2.155.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957.

Cyders, M. A., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity?. *Clinical Psychology Review, 31*(6), 965–982. https://doi.org/10.1016/j.cpr.2011.06.001.

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated?. *Trends in Cognitive Sciences, 24*(4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007.

de Ridder, D. D., de Boer, B. J., Lugtig, P., Bakker, A. B., & van Hooft, E. J. (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. https://doi.org/10.1016/j.paid.2011.01.015.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750.

Dreves, P. A., & Blackhart, G. C. (2019). Thinking into the future: How a future time perspective improves self-control. *Personality and Individual Differences, 149*, 141–151. https://doi.org/10.1016/j.paid.2019.05.049.

Duckworth, A., & Kern, M. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality, 45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004.

Duckworth, A. L., Tsukayama, E., & Kirby, T. A. (2013). Is it really self-control? Examining the predictive power of the delay of gratification task. *Personality and Social Psychology Bulletin, 39*(7), 843–855. https://doi.org/10.1177/0146167213482589.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143–149. https://doi.org/10.3758/BF03203267.

Fine, A., Steinberg, L., Frick, P. J., & Cauffman, E. (2016). Self-control assessments and implications for predicting adolescent offending. *Journal of Youth and Adolescence, 45*(4), 701–712. https://doi.org/10.1007/s10964-016-0425-2.

Frye, C. C., Galizio, A., Friedel, J. E., DeHart, W. B., & Odum, A. L. (2016). Measuring delay discounting in humans using an adjusting amount task. *Journal of Visual Experiments, 107*. https://doi.org/10.3791/53584 e53584.

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and Social Psychology Review, 15*(4), 352–366. https://doi.org/10.1177/1088868311411165.

Godin, G., & Shephard, R. J. (1985). A simple method to assess exercise behavior in the community. *Canadian Journal of Applied Sport Sciences, 10*(3), 141–146. https://doi.org/10.2466/03.27.PMS.120v19x7.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493.

Gray, J. C., Amlung, M. T., Palmer, A. A., & MacKillop, J. (2016). Syntax for calculation of discounting indices from the Monetary choice questionnaire and probability discounting questionnaire. *Journal of The Experimental Analysis of Behavior, 106*(2), 156–163. https://doi.org/10.1002/jeab.221.

Güre, İ. (2012). Understanding consumers’ impulse buying behavior (Doctoral dissertation).

Hagger, M. S., Panetta, G., Leung, C. M., Wong, G. G., Wang, J. C., Chan, D. K., ... Chatzisarantis, N. L. (2013). Chronic inhibition, self-control and eating behavior: Test of a ‘resource depletion’ model. *PLoS ONE, 8*(10). https://doi.org/10.1371/journal.pone.0076888 e76888.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136*(4), 495–525. https://doi.org/10.1037/a0019486.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36. https://doi.org/10.1007/BF02291575.

Kirby, K. N., Petry, N. M., & Bickel, W. K. (1999). Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of Experimental Psychology: General, 128*(1), 78–87. https://doi.org/10.1037/0096-3445.128.1.78.

Koffarnus, M. N., & Bickel, W. K. (2014). A 5-trial adjusting delay discounting task: Accurate discount rates in less than one minute. *Experimental and Clinical Psychopharmacology, 22*(3), 222–228. https://doi.org/10.1037/a0035973.

Lane, S. D., Cherek, D. R., Rhodes, H. M., Pietras, C. J., & Tcheremissine, O. V. (2003). Relationships among laboratory and psychometric measures of impulsivity: Implications in substance abuse and dependence. *Addictive Disorders & Their Treatment, 2*(2), 33–40. https://doi.org/10.1097/00132576-200302020-00001.

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

Magen, Eran, & Gross, James J. (2010). The cybernetic process model of self-control: Situation- and person-specific considerations. In Rick H. Hoyle (Ed.), *Handbook of personality and self-regulation* (pp. 353–374). Oxford, UK: Wiley-Blackwell. https://doi.org/10.1002/9781444318111.ch16.

McDonald, J. H. (2014). *Handbook of biological statistics* (3rd ed.). Baltimore, Maryland: Sparky House Publishing. Retrieved from http://www.biostathandbook.com/multiplecomparisons.html.

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00169.

Mischel, W. (1958). Preference for delayed reinforcement: An experimental study of a cultural observation. *The Journal of Abnormal and Social Psychology, 56*(1), 57–61. https://doi.org/10.1037/h0041895.

Mischel, W. (1961). Preference for delayed reinforcement and social responsibility. *The Journal of Abnormal and Social Psychology, 62*(1), 1–7. https://doi.org/10.1037/h0048263.

Mischel, W., & Ebbesen, E. B. (1970). Attention in delay of gratification. *Journal of Personality and Social Psychology, 16*(2), 329–337. https://doi.org/10.1037/h0029815.

Mischel, W., Ebbesen, E. B., & Raskoff Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology, 21*(2), 204–218. https://doi.org/10.1037/h0032198.

Mischel, W., Shoda, Y., & Peake, P. K. (1988). The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality and Social Psychology, 54*(4), 687–696. https://doi.org/10.1037/0022-3514.54.4.687.

Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science, 244*(4907), 933–938. https://doi.org/10.1126/science.2658056.

Mitchell, L. A., MacDonald, R. A., & Brodie, E. E. (2004). Temperature and the cold pressor test. *The Journal of Pain, 5*(4), 233–237. https://doi.org/10.1016/j.jpain.2004.03.004.

Monterosso, J., & Luo, S. (2013). Willpower is not synonymous with 'executive function'. *Behavioral and Brain Sciences, 36*(6), 700–701. https://doi.org/10.1017/S0140525X1300112X.

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as a limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology, 74*(3), 774–789. https://doi.org/10.1037/0022-3514.74.3.774.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19*(6), 625–666. https://doi.org/10.1521/soco.19.6.625.20886.

Patton, J., Stanford, M. S., & Barratt, E. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology, 51*(6), 768–774. https://doi.org/10.1002/1097-4679.

Peake, P. K., Hebl, M., & Mischel, W. (2002). Strategic attention deployment for delay of gratification in working and waiting situations. *Developmental Psychology, 38*(2), 313–326. https://doi.org/10.1037/0012-1649.38.2.313.

Peckerman, A., Saab, P. G., Llabre, M. M., Hurwitz, B. E., McCabe, P. M., & Schneiderman, N. (1998). Cardiovascular and perceptual effects of reporting pain during the foot and forehead cold pressor tests. *International Journal of Behavioral Medicine, 5*(2), 106–117. https://doi.org/10.1207/s15327558ijbm0502_2.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*(1), 3–25. https://doi.org/10.1080/00335558008248231.

Reynolds, B., Ortengren, A., Richards, J. B., & de Wit, H. (2006). Dimensions of impulsive behavior: Personality and behavioral measures. *Personality and Individual Differences, 40*(2), 305–315. https://doi.org/10.1016/j.paid.2005.03.024.

Reyna, V. F., & Wilhelms, E. A. (2017). The gist of delay of gratification: Understanding and predicting problem behaviors. *Journal of Behavioral Decision Making, 30*(2), 610–625. https://doi.org/10.1002/bdm.1977.

Reynolds, B., Richards, J. B., Horn, K., & Karraker, K. (2004). Delay discounting and probability discounting as related to cigarette smoking status in adults. *Behavioural Processes, 65*(1), 35–42. https://doi.org/10.1016/S0376-6357(03)00109-8.

Richards, J. B., Zhang, L., Mitchell, S. H., & de Wit, H. (1999). Delay or probability discounting in a model of impulsive behavior: Effect of alcohol. *Journal of The Experimental Analysis of Behavior, 71*(2), 121–143. https://doi.org/10.1901/jeab.1999.71-121.

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption: II. *Addiction, 88*(6), 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x.

Saunders, B., Milyavskaya, M., Etz, A., Randles, D., & Inzlicht, M. (2017). Reported self-control does not meaningfully assess the ability to override impulses. Retrieved from osf.io/8etus. doi: 10.17605/osf.io/bxfsu

Schmeichel, B. J., & Zell, A. (2007). Trait self-control predicts performance on behavioral tests of self-control. *Journal of Personality, 75*(4), 743–755. https://doi.org/10.1111/j.1467-6494.2007.00455.x.

Sethi, A., Mischel, W., Aber, J. L., Shoda, Y., & Rodriguez, M. L. (2000). The role of strategic attention deployment in development of self-regulation: Predicting preschoolers' delay of gratification from mother–toddler interactions. *Developmental Psychology, 36*(6), 767–777. https://doi.org/10.1037/0012-1649.36.6.767.

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of 'impulsive' behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin, 140*(2), 374–408. https://doi.org/10.1037/a0034418.

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology, 26*(6), 978–986. https://doi.org/10.1037/0012-1649.26.6.978.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24–31. https://doi.org/10.1177/0098628316677643.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651.

Tangney, J., Baumeister, R., & Boone, A. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72*(2), 271–322. https://doi.org/10.1111/j.0022-3506.2004.00263.x.

van der Maas, H. L., Kan, K. J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. *Seriously. Journal of Intelligence, 2*(1), 12–15. https://doi.org/10.3390/jintelligence2010012.

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends In Cognitive Sciences, 12*(11), 418–424. https://doi.org/10.1016/j.tics.2008.07.005.

Watson, S. J., & Milfont, T. L. (2017). A short-term longitudinal examination of the associations between self-control, delay of gratification and temporal considerations. *Personality and Individual Differences, 10657–60*. https://doi.org/10.1016/j.paid.2016.10.023.

White, J. L., Moffitt, T. E., Caspi, A., Bartusch, D. J., Needles, D. J., & Stouthamer-Loeber, M. (1994). Measuring impulsivity and examining its relationship to delinquency. *Journal of Abnormal Psychology, 103*(2), 192–205. https://doi.org/10.1037/0021-843X.103.2.192.

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30*(4), 669–689. https://doi.org/10.1016/S0191-8869(00)00064-7.

Xu, S., Fang, Z., & Rao, H. (2013). Real or hypothetical monetary rewards modulates risk taking behavior. *Acta Psychologica Sinica, 45*(8), 874–886. https://doi.org/10.3724/SP.J.1041.2013.00874.