# A large-scale analysis of test-retest reliabilities of self-regulation measures

Running head: Retest reliabilities of self-regulation measures

A. Zeynep Enkavi[1], Ian W. Eisenberg[1], Patrick G. Bissett[1], Gina L. Mazza[2], David P. MacKinnon[3], Lisa A. Marsch[4], Russell A. Poldrack[1]
[1]Department of Psychology, Stanford University, Stanford, CA 94305
[2]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ 85259
[3]Department of Psychology, Arizona State University, Tempe, AZ 85281
[4] Geisel School of Medicine, Dartmouth College, Hanover, NH 03755

Author Note

These data have previously been presented as a poster at the Society for Neuroeconomics Annual meeting in Toronto (2017) and Society for Judgment and Decision Making Annual meeting in Vancouver (2017).

Correspondence concerning this article should be addressed to A. Zeynep Enkavi, Department of Psychology, Stanford University, Stanford, CA 94305.
Phone: 215-760-9290
Email: zenkavi@stanford.edu
Classification: Social Sciences > Psychological and Cognitive Sciences

# Abstract

The ability to regulate behavior in service of long-term goals is a widely studied psychological construct known as self-regulation. This wide interest is in part due to the putative relations between self-regulation and a range of real-world behaviors. Self-regulation is generally viewed as a trait, and individual differences are quantified using a diverse set of measures including self-report surveys and behavioral tasks. Accurate characterization of individual differences requires measurement reliability, a property frequently characterized in self-report surveys, but rarely assessed in behavioral tasks. We remedy this gap by (1) providing a comprehensive literature review on an extensive set of self-regulation measures, and (2) empirically evaluating retest reliability in this battery of measures in a new sample. We find that self-report survey measures of self-regulation have high test-retest reliability while measures derived from behavioral tasks do not. This holds both in the literature and in our sample, though the reliability estimates in the literature are highly variable. We confirm that this is due to differences in between-subjects variability. We also compare different types of task measures (e.g., model parameters vs. raw response times) in their suitability as individual difference measures, finding that certain model parameters are as stable as raw measures. Our results provide greater psychometric footing for the study of self-regulation and provide guidance for future studies of individual differences in this domain.

Keywords: self-regulation, retest reliability, individual differences

# Significance Statement

Self-regulation is a psychological construct that is characterized using a broad set of measures and is thought to be related to a number of real-world outcomes. However, the reliability of

many of these measures is unclear. This paper reviews the literature on reliability of self-regulation measures, and characterizes long-term retest reliability in a large sample of individuals completing an extensive battery of measures. The results show that while self-report measures have generally high reliability, behavioral task measures have substantially lower reliability, raising questions about their ability to serve as trait-like measures of individual differences.

# Author Contributions

LAM and RAP designed research, AZE and IWE performed research, AZE analyzed data, AZE, IWE, PGB, GLM, DPM, RAP wrote the paper

# Introduction

The ability to control behavior in service of goals, known as *self-regulation*, is a fundamental aspect of adaptive behavior and central to theories in nearly every area of psychology. Individual differences in self-regulatory ability are thought to be associated with a number of maladaptive behaviors in the real world, including drug abuse (1, 2), problem gambling (3–6), and overeating (7–9). Self-regulation is also thought to play a critical role in behavior change, bolstering the individual against temptations to revert to older behaviors (1, 10, 11), though its role as a moderator of behavior change has recently been challenged (12). Self-regulation, when conceptualized as a personality trait, has generally been measured using self-report surveys that focus on various aspects of naturalistic behavior including impulsivity, sensation-seeking, goal-directedness, and risk-taking.

A central challenge for psychological science is to identify the psychological mechanisms that underlie self-regulatory functions. For example, behavioral tasks, often involving speeded choice responses, are commonly used to compare conditions and isolate component processes. Within cognitive psychology and cognitive neuroscience, there has been particular interest in isolating mechanisms involved in "cognitive control" (13, 14). Candidate mechanisms include the ability to interrupt or preempt a particular behavior (know as *response inhibition*), the ability to rapidly switch between behavioral or task sets (known as *set shifting* or *switching*), and the ability to resist interference from irrelevant information (known as *resistance to distractor interference*). Similarly, researchers in the domain of decision making have focused on the ability to delay gratification in service of larger rewards in the future (known as *delay discounting*), which is thought to relate to a number of real world outcomes (2, 15–17). Given that tasks are intended to capture the mechanisms underlying self-regulation, they would be expected to relate to self-report surveys of self-regulation, but the evidence is mixed. (18–20).

One potential complicating factor in assessing the relation between behavioral task performance and self-report measures is that their psychometric features may differ. In particular, whereas the assessment of retest reliability is a nearly ubiquitous aspect of the development of survey measures, it is rarely assessed in the development of novel behavioral tasks. Further, when assessed in behavioral tasks, it has often been found to fall far short of the common criterion of 75% (21–23). Therefore, it is difficult to determine whether the weak relationship between different measures of self-regulation result from flawed theories or flawed operationalizations of self-regulation.

Here we report a large-scale examination of retest reliability across a broad set of self-report and behavioral task measures relevant to self-regulation and related psychological constructs. We collected retest data on a large battery of measures from 150 participants. These participants comprised a subset of a larger sample acquired in order to model the ontological structure of self-regulation (see 20, 24). We bolstered our dataset with an extensive analysis of the relevant literature for each measure. This allowed us to both compare our data to the literature and assess the relative reliability of data acquired online compared to in-lab samples. Although previous work suggested that data acquired online can exhibit high reliability (25–29), it has not encompassed the breadth of measures relevant to self-regulation collected here. Additionally, the use of a relatively long retest delay (2-4 months) placed the work on the timescale of many studies of behavioral change, providing information relevant to the stability of pre-post intervention comparisons of self-regulatory function. Moreover, using the raw data allowed us to characterize the underlying causes of systematic differences between measure types by isolating the sources of variance for each measure.

With our new dataset we first compared differences between measure modalities (surveys vs. tasks) and recapitulated effects we found in the literature. Then we expanded our analyses to novel comparisons. For example, we compared relative reliability of performance measures quantified using raw variables versus model-based decompositions. We fit the drift-diffusion model (DDM), which transforms raw reaction times and accuracies to the more interpretable latent variables of drift rate (processing speed), threshold (caution that captures speed-accuracy tradeoffs), and non-decision time (perceptual and response execution process).

Another dimension of interest for the behavioral task measures was whether contrast measures (subtraction of one condition from another) intended to isolate putative cognitive processes also served as good trait measures. This subtraction logic is a common strategy when using behavioral tasks, both for raw measures and model parameters. Yet, subtraction of random variables mathematically implies an increase in the contrast measures' error variance and therefore lower retest reliability. We empirically assessed the severity of this decreased reliability for common task contrasts.

By combining an analysis of the literature with a new large dataset involving the largest battery of self-regulation measures to date, we provide a comprehensive picture of the stability of measures of self-regulation.

# Results

## Analysis of prior literature

Our literature review contained 171 dependent measures, 154 papers, 17550 participants and 583 data points on retest reliability (Fig. 1). We first tested for systematic differences between dependent measures from tasks and surveys in the literature. Studies reporting retest reliability for surveys had on average 48 more subjects than those reporting retest reliabilities for tasks (95% credible interval for difference = [28, 70]). We then examined whether sample size

and retest delay (see Methods) were associated with the retest reliability of a measure. Using a model including the sample size along with an indicator variable for task vs. survey measures, we found that task measures' reliability estimates were on average 0.139 lower compared to survey measures' (95 % credible interval of difference = [-0.192, -0.088]; mean retest reliability for task measures in the literature = 0.610, for survey measures = 0.716) and that retest reliability decreased by 0.0001 for every additional participant in a study (95% credible interval for decrease = [-0.0002, -0.00001]). To our knowledge, this is the first documentation of such a bias in the literature with respect to reliability measures,which may reflect publication bias and/or variation in undocumented decisions taken by researchers, as discussed further below.

**Fig. 1.** Summary of the literature review for tasks (left) and surveys (right). Each point represents a study containing test-retest reliability data on an unspecified dependent measure for a given task. The size of the point depends on the sample size of the study and the shape depends on the metric that was used to estimate reliability. Each vertical red line indicates 0 reliability.


Adaptive N-back
Angling Risk Always Sunny
Attention Network
Bickel Titrator
Choice Reaction Time
Dietary Decision
Digit Span
Dot Pattern Expectancy
Go/No-go
Holt Laury
Information Sampling
Keep Track
Kirby
Local Global Letter
Probabilistic Selection
Psychological Refractory Period
Raven's
Recent Probes
Shape Matching
Shift
Simon
Simple Reaction Time
Spatial Span
Stop Signal
Stroop
Tower Of London
-0.25
0.00
0.25
0.50
0.75
1.00
Retest Reliability
BIS-BAS
BIS-11
Brief Self Control
Dickman
DOSPOERT RP
DOSPERT RT
Eating
ERQ
Five Facet Mindfulness
Future Time Perspective
Grit Scale
Impulsive Venture
Leisure Time Activity
Mindful Attention Awareness
Selection Optimization Compensation
Sensation Seeking
Ten Item Personality
Theories Of Willpower
Time Perspective
UPPS-P
-0.25
0.00
0.25
0.50
0.75
1.00
Retest Reliability
Sample Size
10
100
1000
Type
ICC
Kendall's Tau
Pearson
Spearman

# Analysis of new dataset

## Data Quality Checks

To ensure data quality we conducted three tests that are described in more detailed in SI: We checked the reliability of the demographic items in our battery, the effect of retest delay on change of subject scores and the correlation between similar survey items. None of our analyses raised concerns and overall provided some degree of assurance that the participants were real people and not automated machines (which is a concern given that the data were collected using Amazon Mechanical Turk).

## Survey and behavioral task reliability in new data

We calculated 372 dependent measures for behavioral tasks and 74 for surveys. Retest reliabilities for each measure were estimated using a nonparametric bootstrap (1000 samples); statistics on these bootstrapped estimates are reported instead of point estimates. We report ICC(3,k) as the main metric of retest reliability, based on its ability to account for various sources of variance separately as outlined in the Methods[*]. The ICC, which ranges from -1 to 1, is a preferred metric for retest reliability and is not biased by sample size (30). Larger values mean that the two scores of a subject for a given measure are more similar to each other than they are to the scores of other subjects. None of our conclusions change using other reliability metrics. The correlation between point estimates of the different reliability metrics for each measure ranged from 0.932 to 0.980 (see Fig. S3 for scatter plots of different reliability metrics).

Mirroring the results in the literature, the average behavioral task measure reliability was 0.391 lower than the average survey measure reliability (95% credible interval for difference = [-0.451, -0.332]). While survey measures had a median ICC of 0.886 (first quartile 0.708, third

[*] In the remainder of this article we will use $r$ to denote Pearson's correlation, $\rho$ to denote Spearman correlation, *ICC* to denote intraclass correlation and $\tau$ to denote Kendall's correlation.

quartile 0.958), cognitive measures had a median of 0.544 (first quartile -0.140, third quartile 0.843).

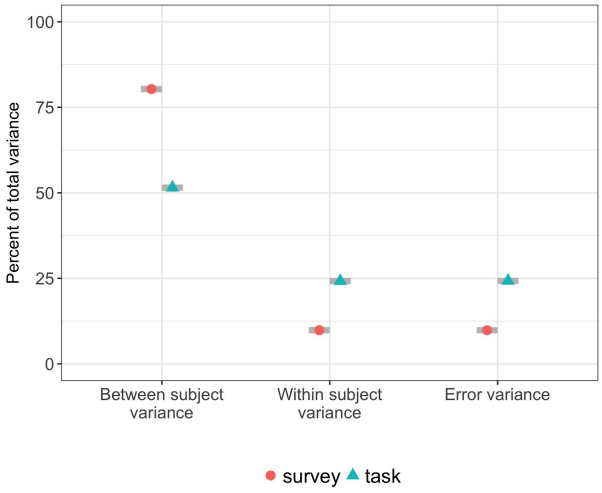
**Fig. 2.** Summary of bootstrapped reliability estimates for tasks (left) and surveys (right). We sampled 150 subjects with replacement 1000 times to create a distribution of bootstrapped reliability estimates for each measure. Reliability measure depicted is ICC. Each vertical red line indicates 0 reliability. Columns following each graph present the mean number of trials used for dependent measures in that task and the number of 'meaningful measures,' dependent measures used in the literature from each task.


-0.25 0.00 0.25 0.50 0.75 1.00
Mean Num Trials
Num Measures
Adaptive N-back 380 3
Angling Risk Always Sunny 21 2
Attention Network 49 10
Bickel Titrator 30 4
Choice Reaction Time 150 2
Cognitive Reflection 6 1
Columbia Card Cold 24 5
Columbia Card Hot 24 5
Dietary Decision 48 1
Digit Span 10 2
Directed Forgetting 16 2
Discount Titrate 36 1
Dot Pattern Expectancy 33 13
Go/No-go 175 5
Hierarchical Rule 360 1
Holt Laury 10 1
Information Sampling 10 8
Keep Track 9 1
Kirby 14 8
Local Global Letter 32 21
Motor Selective Stop Signal 90 1
Probabilistic Selection 48 6
Psychological Refractory Period 140 2
Raven's 18 1
Recent Probes 42 4
Shape Matching 139 1
Shift 401 9
Simon 54 9
Simple Reaction Time 149 1
Spatial Span 10 2
Stim Selective Stop Signal 240 1
Stop Signal 304 10
Stroop 63 9
Task Switching 29 2
Tower Of London 12 3
Two Stage Decision 198 2
-0.25 0.00 0.25 0.50 0.75 1.00
Mean Num Trials
Num Measures
BIS-BAS 15 4
BIS-11 7 5
Brief Self Control 13 1
Dickman 12 2
DOSPERT EB 6 5
DOSPOERT RP 6 5
DOSPERT RT 6 5
Eating 9 4
ERQ 5 2
Five Facet Mindfulness 13 6
Future Time Perspective 10 1
Grit Scale 8 1
Impulsive Venture 17 2
Leisure Time Activity 1 1
Mindful Attention Awareness 15 1
Mpq Control 24 1
Selection Optimization Compensation 12 4
Self Regulation 31 1
Sensation Seeking 16 5
Ten Item Personality 2 5
Theories Of Willpower 12 1
Time Perspective 11 5
UPPS-P 12 5

A quantitative explanation for the difference in reliability estimates between surveys and tasks, as recently detailed by Hedge et al. (2017), lies in the difference in sources of variance between these measures. Specifically, the ICC is calculated as the ratio of between-subjects variance versus total variance. Intuitively, measures with high between-subjects variance are better suited for individual difference analyses as they will be sensitive to the differences between the subjects in a sample. Conversely, as Hedge et al. note, behavioral tasks are generally selected on the basis of reliable group effects, which systematically selects for measures with low between-subject variance.

We find that on average 83.36% of survey measures' variance is due to between subjects variability compared to 52.55% of behavioral task measures' (95% credible interval of difference = [25.2, 32.1]; Fig. 5). Conversely, 18.07% of behavioral tasks variance is explained by within-subject variance compared to 5.38% of survey measures (systematic differences between sessions; 95% credible interval of difference = [10.55, 18.36]) and 22.82% by residual variance compared to 8.976% for survey measures (95% credible interval difference = [12.29, 16.69]; model includes random effects for each measure).

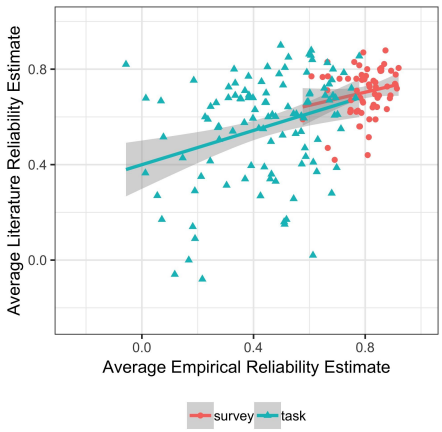
**Fig. 3.** Percentage of variance explained by each of the three sources of variance: between subjects, within subjects (between sessions) and error variance for 1000 bootstrapped samples. Error bars are 95% confidence intervals. The confidence intervals for the bootstrapped samples are quiet small and thus appear like lines compared to the larger markers depicting the means.




## Comparison of literature and new data

To compare our findings to the literature, we first sampled the same number of estimates from our bootstrapped results as we found in the literature for each measure and calculated the correlation between the sampled empirical (i.e. from our data) reliability estimates to those found in the literature. Repeating this 100 times we found that the mean correlation (Fig 2) between our empirical reliability estimates and those based on the literature was 0.274 for behavioral task measures (range = 0.235 - 0.323) and 0.138 for survey measures (range = 0.038 - 0.249).

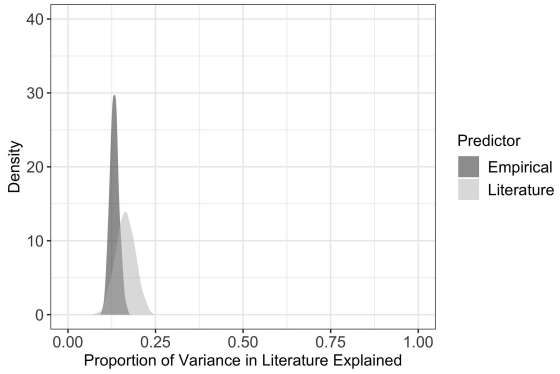
**Fig. 4.** Correlation between mean reliability estimates for each measure found in the literature with the mean reliability from our data.



While the relationship between the empirical and literature-derived reliability estimates seems weak, this must be contextualized by evaluating the variability of retest reliability estimates in the literature. If individual studies in the literature have similarly weak relationships to the literature-wide retest reliability for a given measure (i.e. if the variance of the reliability estimates reported in the literature for a given measure is large), this suggests a general issue of variability in retest reliability estimates across samples and not a specific issue with our sample. Therefore, we compared two types of models: (1) One where we predicted the literature retest reliability using an estimate sampled from the literature review. (2) Another where we predicted the literature retest reliability using the estimate from the new data we collected.

Fig. 3 shows that models using an estimate from the literature to predict the remaining reliability estimates from the literature are systematically better than models using the estimate from our sample. However, the decrease in variance explained using our data is only 2.95% (95% credible interval of difference = [2.78% - 3.13%]) on average, suggesting that published estimates of retest reliability in this domain are relatively noisy. This also suggests that estimating reliability using an online sample does not change conclusions compared to in-lab samples.

**Fig. 5.** Noise ceiling for comparing empirical retest reliability estimates to reliability estimates from the literature. Data come from sampling a single reliability estimate from the literature and using that as a predictor of the remaining retest reliability estimates from the literature versus using the mean estimate from our empirical results as a predictor. Models account for the effect of sample size in the literature and whether the measure is a task or survey variables. The literature samples are significantly better at predicting the rest of the literature than our empirical averages, but the distributions of variance explained across samples are highly overlapping.




## Effect of task length on stability

To compare potential effects of task-specific attributes on retest reliability across tasks, we examined the relationship between the number of trials a task included and its stability. Across non-DDM[†] measures, there was an insignificant 0.023 point change increase in reliability for each additional trial[‡] (95% credible interval of increase = [-0.011, 0.061]).

For tasks for which dependent measures are estimated using many trials one can ask whether the same measure becomes less reliable if fewer trials are used to estimate its reliability. Such analyses would provide a detailed examination of how to extract the most reliable individual difference measure from tasks with measures that have low retest reliabilities, and would address the concern that reliability might be underestimated in the present data due to insufficient numbers of trials. It could also guide researchers in choosing number of trials for long tasks in an informed manner. We provide an example of this approach in the SI; given the open access nature of the data, investigators interested in other tasks can perform similar analyses on those.

---

[†] A detailed analysis of the DDM parameter estimates will be reported elsewhere.

[‡] For measures that were calculated using different numbers of trials for each subject due to timing out or other exclusions we took the mean number of trials used for the measure across all subjects.

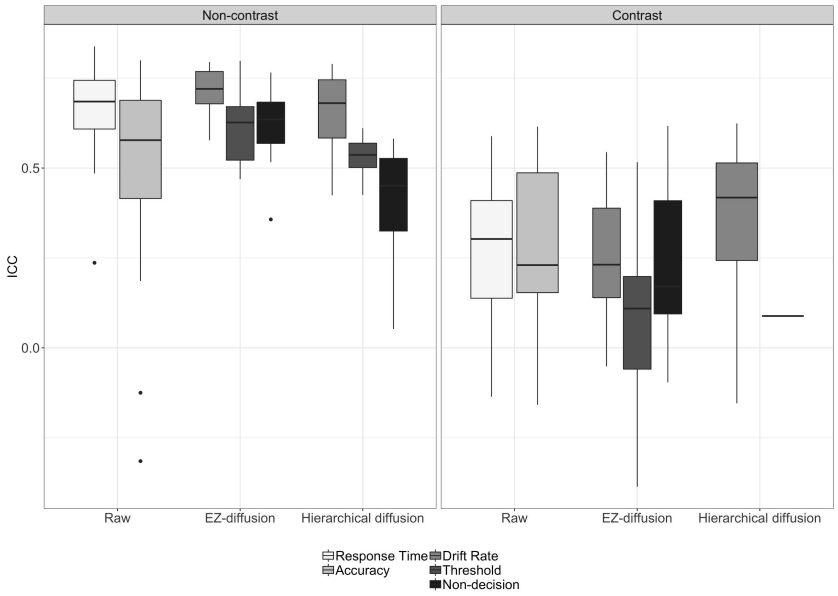
## Comparison of task measure types

Data from any given behavioral task can often be analyzed in various ways, yielding different types of dependent measures. We compare raw measures of response times and accuracies to parameters of the drift diffusion model, which addresses speed-accuracy tradeoffs and offers more interpretable latent variables from a rich literature in mathematical and experimental psychology (31, 32, 33). There are different fitting procedures within this class of models. We chose two approaches to DDM: EZ-diffusion (34) and hierarchical drift diffusion model (HDDM) (35).

The EZ-diffusion method is a set of closed-form expressions that transform mean RT, variability of RT, and accuracy to drift rate, threshold, and non-decision time. The HDDM uses hierarchical Bayesian modeling to allow simultaneous estimation of both group and individual subject parameters. We compare the reliability estimates of raw measures (response times and accuracies) to DDM parameter estimates. Both types of measures can also be 'contrast' and 'non-contrast.'

We found that estimates of non-contrast measures (Fig. 6) are on average 0.356 points more reliable than estimates of contrast measures (95% credible interval of difference = [0.311, 0.406]). This is not surprising given the summing of the variance in the difference score. Of concern, however, is the fact that contrast measures had low to no reliability (mean = 0.245, SD = 0.242) compared to the moderate reliability of the non-contrast measures (mean = 0.601, SD = 0.183). This is particularly alarming given their common use in cognitive psychology as putative trait measures of cognitive constructs and predictors of real-world outcomes.

**Fig. 6.** Average bootstrapped reliability estimates per measure comparing raw measures and model parameters as well as contrast and non-contrast measures for task measures.




Although individual measures from tasks are not appropriate for individual difference analyses this does not preclude other ways of using them as trait measures. One way to do this is to use a data-driven approach to integrate them and extract scores that may be more stable though not necessarily more predictive of real-world behaviors. An example of this using the same dataset reported in this paper is detailed in Eisenberg et al. (20): Factor scores computed at both time points using the same linear combination of dependent measures correlated highly with each other for 5 task factors ($M$ = .81, min = .76, max = .86) and 12 survey factors ($M$ = .86, min = .75, max = .95). Yet despite adequate reliability for both task and survey factors, only surveys predicted a significant amount of variance in real-world behaviors out of sample (average $R^2$ = .10) whereas tasks did not, either as factors or as separate dependent measures (average $R^2$ = .02)

## Effect of survey length on stability

Mirroring the task analysis, we examined the relationship between the number of items in a survey and its stability. Each additional item used in the calculation of a subscale was associated with an insignificant 0.001 increase in retest reliability (95% credible interval = [-0.001, 0.004]) though as with tasks, surveys could also be analyzed in more detail using item response theory or more sophisticated models given the open availability of the data.

# Discussion

This report provides a systematic characterization of the reliability of self-report and behavioral task measures of the construct of self-regulation. We first summarized the prior literature on the retest reliability of different types of self-regulation measures. We found that while psychometric studies of survey measures have larger sample sizes than task measures, reliability estimates generally decreased with sample size. On the one hand, this might suggest that smaller studies afford researcher's more control over their measurement, leading to higher reliability. On the other hand, larger sample sizes might be more reflective of the truly lower stability of measures; Hopkins (30) suggests that studies of retest reliability with samples smaller than 50 should be treated as pilot studies for this reason. Studies with smaller samples are more prone to yield variable reliability estimates and coupled with publication bias might inflate the results in the literature. The majority of the results in the literature, particularly those on tasks, have sample sizes in the <50 range. Our new data acquisition had a large sample size and found relatively low reliability of behavioral tasks, consistent with the results from the literature.

Second, we contextualized new results from our battery of self-regulation measures using an extensive literature review. We estimated the general reliability of the literature's own estimates of the retest reliability of self-regulation. This provided a sense of the "noise ceiling" of reliability studies, and a reference point for the expected relationship between any two sets of reliability estimates. Because the literature reliability estimates lacked a strong coherence for many measures, their low correlations with our reliability estimates led to a less than 3% decrease in the predictability of prior literature, suggesting that the results reported on the new dataset here are not far outside what one would expect from the literature.

## Systematic differences in the reliability of self-regulation measures

Literature analysis and our data show that measures of self-regulation that are based on self-report surveys have higher retest reliability than behavioral task measures because of higher between subject variance for survey measures compared to task measures. This suggests that survey measures are more appropriate as trait-level measures suitable for individual difference analyses. Exploratory analyses on task measures suggested that the reliability of DDM parameters did not significantly differ from the reliability of raw measures like response times and accuracy. Researchers may therefore prefer drift diffusion measures given their interpretability.

Revisiting a longstanding question on the reliability of contrast scores, we confirm that they are less reliable than their components in this study. Measures of differences in response times between conditions have lower reliabilities due to correlations between the two measures used in the creation of the difference score (36, 37) and the increase in the variance through subtraction. The novel and concerning point of this finding is that many cognitive measures of interest in the self-regulation literature are contrast measures that have low to no reliability.

## Implications of low reliability for behavioral task measures

Although our conclusion that task measures of self-regulation are less suitable for individual difference analyses might be disappointing, especially in the face of many lines of work showing correlations between these measures and problematic real-world behaviors, it should not be surprising. As Hedge and colleagues (38) argue, behavioral tasks designed with the subtraction logic to isolate specific cognitive processes become well-established in the literature precisely for their low between-subject variability, which necessitates that they will have low retest reliability. For example, one might repeatedly find a significant Stroop effect (difference in the response times between the congruent and incongruent conditions) in samples

measured multiple times, even while the relative distribution of individual response times for the subjects differ. In other words, the task might have low between subject variability and high within subject (between session) variability resulting in low test-retest reliability. This does not invalidate the existence of the Stroop effect but does undermine its suitability as a trait measure. Detailed analyses of sources of variance (within versus between subjects) provides researchers with a priori hypotheses on which measures to expect significant changes in different experimental designs. MacLeod et al. (39) provide an example where they hypothesize that one of three attentional networks from the ANT task is best suited for detecting significant changes in within-subjects designs due to its low within-subject variance but least suited for detecting significant changes in between-subject designs due to its high between-subject variance. Furthermore, task measures can be integrated using data-driven approaches to extract factor scores that are more stable and potentially more suitable for trait-like treatment. Using this approach, we indeed found more stable dependent measures, though they were not more predictive of real-world behaviors (20).

On the other hand, different psychometric properties of measures serve different purposes. For example, while high retest reliability is desirable for measures that will be used in trait-like characteristic analyses, it is neither a necessary nor a sufficient condition for the responsiveness of a measure to capture change over time (40). Although our results provide practical guidelines for researchers interested in these measures they do not answer how these measures relate to the construct of "self-regulation." While the retest reliability of a measure has consequences on the limits of its correlation with other measures, specifically for any two variables the correlation between them must be smaller than the square root of the reliability of

each measure (36, 41, 42), the question of validity remains a separate one that we address in related work (20).

## Conclusions

Self-regulation is a central construct in many theories of behavior and is often targeted by interventions to reduce or control problem behavior. Our study of self-report and task measures of self-regulation suggests stability in many self-report measures and less stability in behavioral task measures. We hope that these analyses and open data provide guidance for future individual difference work in self-regulation.

# Materials & Methods

## Sample

Participants in this study were a subset from a larger study of self-regulation (24) conducted on Amazon Mechanical Turk (MTurk). Invitations were sent to 242 of 522 participants (52% female, age: mean = 34.1, median = 33, range = 21-60) who had satisfactorily completed the first wave of data collection between July and September 2016. The final sample for the retest study consisted of 150 participants (52.7% female, age: mean = 34.5, median = 33, range = 21-60) whose data passed basic quality checks as described in Table S1. The sample size was specified prior to data collection based on financial constraints. Instead of inviting all 522 eligible participants at once we invited randomly selected subsets of participants in small batches. This addressed preferentially sampling the most motivated, prompt subjects who may systematically differ from the full sample. Each batch was given a week to complete the battery. Data collection for the second wave took place between November 2016 and March 2017. The mean number of days between the two waves was 111 days (median = 115 days ; range = 60-228 days). Of the 242 participants invited 175 participants started the battery and 157 completed the battery. This study was approved by the Stanford Institutional Review Board (protocol IRB-34926).

The data collection platform as well as the details of the data analysis pipeline including links to analysis scripts and interactive visualizations are listed in the SI.

## Literature review

The literature review was conducted on Google Scholar, which was chosen for its breadth. Our strategy consisted of the following steps: (1) Manually check the reference article for a given task or survey (i.e. the article that described the task or survey for the first time) for retest reliability data. (2) Search within the full text of the articles that cite the reference article

for the term ‘retest.’ (3) Examine each of these resulting articles up to the first 100 results ordered by the number of times they have been cited. (4) Scan the abstract and the methods sections of each article to determine whether the article reports original empirical retest results. (5) Extract the empirical results from the Results section making sure to include the following information: (a) the type of retest reliability statistic, (b) the magnitude of the statistic, (c) the dependent measure the retest reliability data pertains to, (d) delay between the two measurements, (e) sample size, (f) any differences from the procedure used in our battery, and (g) Article reference. (6) If the resulting article cites other articles with retest reliability for the measure, then find and examine them for retest reliability data using the same method as above. (7) If the reference article describes a version of the task that is modified for specific purposes (e.g. the Shift Task is a modified version of the older Wisconsin Card Sorting task) then find the reference article for the parent task and apply the same search routine for retest reliability on the parent task.

Detailed descriptions of all behavioral tasks and surveys as well as the findings on retest reliability for all of them are listed in the SI.

# Acknowledgements

This work was supported by the National Institutes of Health (NIH) Science of Behavior Change Common Fund Program through an award administered by the National Institute for Drug Abuse (NIDA) (UH2DA041713; PIs: Marsch, LA & Poldrack, RA). Additional support was provided by NIDA P30DA029926.

# References


1. Prochaska JO, DiClemente CC, Norcross JC (1992) In search of how people change. Applications to addictive behaviors. *Am Psychol* 47(9):1102–1114.

2. Kirby KN, Petry NM, Bickel WK (1999) Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *J Exp Psychol Gen* 128(1):78.

3. Alessi SM, Petry NM (2003) Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behav Processes* 64(3):345–354.

4. Kertzman S, et al. (2008) Go--no-go performance in pathological gamblers. *Psychiatry Res* 161(1):1–10.

5. Lawrence AJ, Luty J, Bogdan NA, Sahakian BJ, Clark L (2009) Impulsivity and response inhibition in alcohol dependence and problem gambling. *Psychopharmacology* 207(1):163–172.

6. Fuentes D, Tavares H, Artes R, Gorenstein C (2006) Self-reported and neuropsychological measures of impulsivity in pathological gambling. *J Int Neuropsychol Soc* 12(6):907–912.

7. Nederkoorn C, Smulders FTY, Havermans RC, Roefs A, Jansen A (2006) Impulsivity in obese women. *Appetite* 47(2):253–256.

8. Nederkoorn C, Braet C, Van Eijs Y, Tanghe A, Jansen A (2006) Why obese children cannot resist food: the role of impulsivity. *Eat Behav* 7(4):315–322.

9. Hendrickson KL, Rasmussen EB (2013) Effects of mindful eating training on delay and probability discounting for food and money in obese and healthy-weight individuals. *Behav Res Ther* 51(7):399–409.

10. Rozensky RH, Bellack AS (1974) Behavior change and individual differences in self-control. *Behav Res Ther* 12(3):267–268.

11. Bickel WK, Vuchinich RE (2000) *Reframing Health Behavior Change With Behavioral Economics* (Psychology Press).

12. Stautz K, Zupan Z, Field M, Marteau TM (2018) Does self-control modify the impact of interventions to change alcohol, tobacco, and food consumption? A systematic review. *Health Psychol Rev* 12(2):157–178.

13. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.

14. Miyake A, et al. (2000) The Unity and Diversity of Executive Functions and Their Contributions to Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. *Cogn Psychol* 41(1):49–100.

15. Mischel W, Shoda Y, Rodriguez MI (1989) Delay of gratification in children. *Science* 244(4907):933–938.

16. Baker F, Johnson MW, Bickel WK (2003) Delay discounting in current and never-before cigarette smokers: similarities and differences across commodity, sign, and magnitude. *J Abnorm Psychol* 112(3):382.

17. Meier S, Sprenger CD (2012) Time discounting predicts creditworthiness. *Psychol Sci* 23(1):56–58.

18. Duckworth AL, Kern ML (2011) A Meta-Analysis of the Convergent Validity of Self-Control Measures. *J Res Pers* 45(3):259–268.

19. Nęcka E, Gruszka A, Orzechowski J, Nowak M, Wójcik N (2018) The (In)significance of Executive Functions for the Trait of Self-Control: A Psychometric Study. *Front Psychol* 9:1139.

20. Eisenberg IW, et al. (2018) Uncovering mental structure through data-driven ontology discovery. Available at: https://psyarxiv.com/fvqej/.

21. Cicchetti DV, Sparrow SA (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic* 86(2):127–137.

22. Fleiss JL, Levin B, Paik MC (2013) *Statistical Methods for Rates and Proportions* (John Wiley & Sons).

23. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

24. Eisenberg IW, et al. (2017) Applying novel technologies and methods to inform the ontology of self-regulation. *Behav Res Ther*.

25. Paolacci G, Chandler J, Ipeirotis PG (2010) Running Experiments on Amazon Mechanical Turk. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1626226 [Accessed May 3, 2018].

26. Horton JJ, Rand DG, Zeckhauser RJ (2011) The online laboratory: conducting experiments in a real labor market. *Exp Econ* 14(3):399–425.

27. Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspect Psychol Sci* 6(1):3–5.

28. Behrend TS, Sharek DJ, Meade AW, Wiebe EN (2011) The viability of crowdsourcing for survey research. *Behav Res Methods* 43(3):800–813.

29. Crump MJC, McDonnell JV, Gureckis TM (2013) Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8(3):e57410.

30. Hopkins WG (2000) Measures of reliability in sports medicine and science. *Sports Med* 30(1):1–15.

31. Wickelgren WA (1977) Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol* 41(1):67–85.

32. Ratcliff R (1978) A theory of memory retrieval. *Psychol Rev* 85(2):59.

33. Ratcliff R, Smith PL, Brown SD, McKoon G (2016) Diffusion Decision Model: Current Issues and History. *Trends Cogn Sci* 20(4):260–281.

34. Wagenmakers E-J, van der Maas HLJ, Grasman RPPP (2007) An EZ-diffusion model for response time and accuracy. *Psychon Bull Rev* 14(1):3–22.

35. Wiecki TV, Sofer I, Frank MJ (2013) HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform* 7:14.

36. Salthouse TA, Hedden T (2002) Interpreting reaction time measures in between-group comparisons. *J Clin Exp Neuropsychol* 24(7):858–872.

37. Caruso JC (2004) A Comparison of the Reliabilities of Four Types of Difference Scores for Five Cognitive Assessment Batteries. *Eur J Psychol Assess* 20(3):166–171.

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

39. MacLeod JW, et al. (2010) Appraising the ANT: Psychometric and theoretical considerations of the Attention Network Test. *Neuropsychology* 24(5):637.

40. Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 40(2):171–178.

41. Paap KR, Sawi O (2016) The role of test-retest reliability in measuring individual and group differences in executive functioning. *J Neurosci Methods* 274:81–93.

42. Spearman C (1904) The Proof and Measurement of Association between Two Things. *Am J Psychol* 15(1):72–101.

PNAS
www.pnas.org

# Supplementary Information for

## A large-scale analysis of test-retest reliabilities of self-regulation measures

A. Zeynep Enkavi, Ian W. Eisenberg, Patrick G. Bissett, Gina L. Mazza, David P. MacKinnon, Lisa A. Marsch, Russell A. Poldrack

A. Zeynep Enkavi
Email: zenkavi@stanford.edu

**This PDF file includes:**

## Supplementary Information Text

### Measures of retest reliability used for new data

The literature review presents data on the retest reliability (Table S1) of the paradigms included in our battery. Here we present a brief overview of the different types of test-retest reliability statistics that will be listed later. More detailed analyses can be found elsewhere (1–3).

When reporting analyses on data from our sample we chose *ICC* (specifically, ICC(3,k) = (between subjects variance - error variance)/between subjects variance) as the main retest reliability statistic though the results are qualitatively similar under all metrics. In contrast, we used Pearson correlation as the retest reliability metric when any analysis involved the literature-derived reliability estimates, as most results in the literature used this metric. When predicting the literature using our sample, we also averaged the Pearson correlation estimate for each measure across bootstrapped samples.

### Data collection platform

Data were collected using the Experiment Factory (45), an open-source development and deployment platform for behavioral experiments. It consists of a repository of Javascript experiments relying on the jsPsych library (46) and a Python application to run experiments both locally or on platforms such as MTurk. Experiment batteries contain modules for all the surveys and behavioral tasks a participant is asked to complete. When a worker accepts the human intelligence task (“HIT”) for the study on MTurk, their worker ID is associated with a personalized URL that allows them log back in and complete their remaining tasks and surveys. The presentation of the tasks and surveys is randomized for each subject.

### Analyses

Dependent measure calculation including DDM parameter fitting was conducted on the Stanford Sherlock Cluster. An interactive visualization tool presenting results from both the

literature review and raw data analysis at a measure level is available at https://zenkavi.github.io/SRO_Retest_Analyses/output/reports/Interactive_Figures.nb.html The code for dependent measure calculation can be found at: https://github.com/IanEisenberg/expfactory-analysis/tree/master/expanalysis/experiments. The data are available at: https://github.com/IanEisenberg/Self_Regulation_Ontology/tree/master/Data Analyses on the reliability of measures were performed in R (Version 3.3.2) using the packages listed on: https://github.com/zenkavi/SRO_Retest_Analyses/blob/master/output/reports/sessionInfo.txt All analysis code and data are available at https://zenkavi.github.io/SRO_Retest_Analyses/output/reports/SRO_Retest_Analyses.nb.html

**Behavioral Task descriptions**

In this section, we provide brief descriptions of the behavioral tasks included in our battery and summarize the literature on their retest reliabilities. We also list the dependent measure(s) that were selected for each task for use in reliability calculations. The dependent measures for each task were a combination of measures we found in the literature and modeling approaches of interest to us (e.g. DDM). The literature review is summarized at the task level in Fig 1. This figure does not include measures such as DDM parameters, as we did not find reliability estimates for them in the literature. For comparability the reliability estimates from the new dataset that are summarized in Figure 4 also only use these 'meaningful measures' found in the literature. More details on the reliabilities of specific measures can be found in the interactive figures listed in Analyses.

**Adaptive N-back Task**

In this task (4) subjects view a stream of letters on the screen, one letter at a time. They press one button when the letter on the screen matches the letter presented N (load) trials back, where N is specified at the beginning of each block; the case of the letter is irrelevant. They press another button for all other letters. Each block consists of twenty plus the load number of letters. The load is increased if the subject has made fewer than three mistakes in the previous block. It is decreased if the subject has made more than five mistakes. Each subject goes through twenty blocks. The n-back task has been widely used in the literature, with multiple versions that differ in the type of stimulus (e.g. numeric, alphabetic, spatial) as well as its delivery (e.g. visual, auditory). The dependent measures for which reliability was calculated for this task are accuracy, mean response time, mean load across blocks and drift diffusion model (DDM) parameters for all trials.

While a large literature exists on working memory training over several weeks using the N-back task, (see 5 for a meta-analysis), test-retest reliabilities of the multiple measurements during training have rarely been reported. We found no retest reliability data for the adaptive version of the task that we used in our battery, but reliability has been reported for a number of other versions. Hockey and Geffen (6) reported *r*'s ranging from 0.49 to 0.73 for accuracy and from 0.69 to 0.86 for response times depending on the load level for 70 students one week apart. Their task, however used of visuospatial stimuli instead of letters and was not adaptive. Van Leeuwen and colleagues (7) found *r*'s of 0.65 for total number of correct responses in the 2-back condition and 0.70 in the 3-back condition for 105 children ranging between 8 and 11 in age between two measurements separated by two to three weeks. This declined to 0.16 for 2-back but stays the same for the 3-back condition for 29 adolescents (ages 14-20) for the respective conditions. Their task, however, was not adaptive and used stimuli adapted for children. For

another non-adaptive version with visuospatial stimuli, Studer-Luethi et al. (8) reported a *r* of 0.69 for 112 Chinese subjects tested four weeks apart.

**Angling Risk Task**

In this task (9), subjects play a fishing game for thirty rounds for four conditions. A simulated pond contains a population of red fish along with one blue fish; each red fish caught translates to earnings in the round, while catching the blue fish results in loss of all earnings accrued in the round. The subject can end the round at any point, which places the earnings for the round into their permanent earnings bank. In the original task there were two weather conditions, corresponding to the subject's knowledge of the distribution of red/blue fish: In the 'sunny' condition subjects could always see how many fish there were in the lake, in the 'cloudy' condition they could not. Due to time constraints on the total length of our task battery we only used the "sunny" condition. There were also two sampling rules. In the 'keep' condition, caught red fish were not replaced, simulating sampling without replacement and thus increasing the probability of catching a blue fish after each draw. In the 'release' condition the red fish were released so the number of fish in the lake remained constant for the whole round, effectively sampling with replacement and thus maintaining a consistent probability of catching a blue fish. The number of fish varied between 1 and 200 for each round. Total score on this task contributed to the final bonus each participant received. The dependent measures chosen for this task were the mean number clicks for trials where a blue fish was not caught (adjusted clicks), the percentage of trials where a blue fish was caught and the total score for both release rules.

This task is modeled after the balloon analogue risk taking task (BART: 10) where fishing tournaments are replaced with the pumping of a balloon. While we have not found data on the test-retest reliability of ART, the reliability of the BART has been investigated. Same day *r*'s for this measure ranged from 0.62 to 0.82 (11). White, Lejuez and de Wit (12) reported test-

retest reliabilities (*r*'s) ranging from 0.66 to 0.78 (depending on the size of reward with each pump) for the adjusted mean pumps, the putative risk-taking dependent measure, for the related task of Balloon Analogue Risk Task across three days (n=38). Weafer et al. (13) reported test-retest reliability of 0.79 for the same measure (n=119) with a significant increase in risk taking across the two time points (mean delay 8.6 days). For 275 adolescents (ages 9-12 at initial enrollment) tested for at least two waves across two years MacPherson et al. (14) found *r*'s for adjusted pumps using BART ranging from 0.39 to 0.67.

**Attentional Network Task**

In this task (15) subjects indicate the direction of a center arrow that is surrounded by two flankers on each side. The set of five stimuli (target + flankers) can appear below or above a center fixation cross. There are three conditions depending on the direction of the surrounding arrows: *incongruent* if flankers are arrows pointing in the opposite direction than the target stimulus; *congruent* if they are arrows pointing in the same direction; and *neutral* if the flankers are horizontal lines instead of arrows. There are four conditions depending on the cue (a briefly presented star) before the presentation of the target stimuli: "no cue" trials in which no cue is presented before the target stimulus, "double cue" trials in which two simultaneous cues are flashed above and below the fixation cross, "center cue" trials in which the cue is flashed in the location of the fixation cross, and "spatial cue" trials in which the cue is flashed in the location where the target stimulus will follow. Subjects complete 24 practice trials and 144 experimental trials (2 (locations) x 4 (cues) x 2 (direction) x 3 (flanker) x 3 (blocks)). Differences in error rates and mean response times between different conditions provides putative measures of three components (networks) of attention: alerting (double cue - no cue), orienting (informative spatial - central cue) and executive control (congruent - incongruent trials). The dependent measures we chose from this task are overall and attention component specific accuracy, overall response

time and specifically on, congruent, incongruent, neutral trials as well as for each attention component contrast, and overall DDM parameters as well as attention component contrasts.

In their paper describing the development of this task Fan et al. (15) reported *r*'s of 0.52 for alerting, 0.61 for orienting, 0.77 for executive control networks and 0.87 for all the response times for 40 adults with a delay of 10 minutes. In a preliminary study checking for heritability of attentional networks Fan et al. (16) found *r*'s correlations of 0.36 for alerting, 0.41 for orienting and 0.81 for conflict for 104 subjects (26 monozygotic twin pairs and 26 age matched dizygotic twins) with a delay of a couple minutes. Contrary to Fan et al.'s (15) original results, Ishigami and Klein (17) found in a small study that between the first two sessions (out of ten) *r*'s are acceptable only for the executive control network (0.86) but not for the alerting (-0.02) and orienting (0.57) networks with ten subjects who complete the task ten times across varying delays with a mean of 8.6 days. For 68 Danish participants across three sessions that were each a week apart, Habekost, Petersen and Vangkilde (18) reported *r*'s correlations of of 0.84 and 0.74 for the executive control network, 0.14 and 0.35 for the alerting network and 0.58 and 0.59 for the orienting network. They also found strong decreases in the executive control network scores across sessions and a slight decreases in the orienting network scores while none for the alerting network. For 75 participants tested one week apart, Paap and Sawi (19) found test-retest correlations of 0.642 for RTs in neutral baseline trials, 0.879 for incongruent trials, 0.858 for congruent trials, 0.856 for global RT and 0.515 for the incongruent congruent difference.

**Choice Reaction Time**

In this task (20) subjects see either orange or blue squares on the screen for each trial and are instructed to respond using a different button for each stimulus as quickly as possible. Subjects complete twenty practice trials and three blocks of fifty test trials. The dependent measures from this trial are overall accuracy, median response times and DDM parameters.

For 55 undergraduates re-tested 2-4 weeks later, Barrett et al. (21) reported an *r* of 0.6 for mean response times in correct trials. Kennedy et al. (22) reported an 0.94 for both the two and four choice versions. For 21 subjects tested 4 weeks apart during EEG recording, Williams et al. (23) found an *r* of 0.71 for mean response of time. For twenty participants Deary, Liewald and Nissan (24) reported *r*'s of 0.83 for the mean and 0.62 for the standard deviation of response times for a task with four possible responses. The tests were administered back to back. Other studies compared versions of the this task with different number of response options as well. For the difference between a participant's mean and modal response[§], Weafer et al. (13) reported an *r* of 0.38 for 124 participants and a significant increase in this measure across two time points (9 days). Jones et al. (25) reported *ICC*'s of 0.749 for mean response times of 96 subjects tested a day later.

**Cognitive Reflection Task**

In the classical version of the task (26) subjects answer three questions that have numeric answers. The questions are worded such that there is a spontaneous but erroneous answer and the correct answer typically requires slower and more thoughtful responses. Because our sample may have been familiar with the questions of the classical version we have used three lesser known items from each of Toplak, West and Stanovich's (27) and Primi et al.'s (28) expansions.

---

[§] Weafer et al. explain their reasoning behind this measure that is not commonly calculated in the literature as follow: "Based on a participant's distribution of RT's, a deviation from the mode score was calculated as the difference between a participant's mean and modal RT. This value represents the proportion of unusually long RT's, which are inferred to reflect momentary lapses in attention."

We used these same items both for the first and the second testing. The dependent measures from this task are the proportion of correct and intuitive responses.

We have not found data on its test-retest reliability. Some work suggests that studying test-retest reliability for this measure might not be useful because subjects who have seen the questions once perform significantly different than naive subjects (29, 30). As suggested above, different forms have been introduced that could potentially alleviate this problem, but we did not find any evidence of test-retest on two different from of this task.

**Columbia Card Sorting Task**

In this task (31) subjects play a card game with the goal in each round to collect as many points as possible by flipping cards from a deck of 32. Each deck contains gain and loss cards; each gain card is worth either 10 or 30 points, each loss card costs either 250 or 750 points, and there are either 1 or 3 three loss cards in each round. All the round information (the number of win cards in the deck, the win amount per win card, the loss amount per loss card and the number of loss cards in the deck) is always on display throughout the round. Subjects play 24 rounds in two conditions. In the *hot* condition they flip each card individually and see the outcome of the card sequentially whereas in the *cold* condition they only indicate how many cards they would want to flip given the round information. The dependent measures from this task are the mean number of cards chosen in all rounds, information use, gain, loss and probability sensitivity for both the hot and the cold conditions. The sensitivities are operationalized respectively as the coefficients when regressing the number of cards chosen over the gain amount, loss amount and the number of loss cards. Information use is the number of coefficients that are significant from this regression.

We were not able to find any data on the test-retest reliability of this specific sequential risk taking task.

**Delay Discounting Tasks**

**Adaptive Adjusting Amount Delay Discounting Task:**

In this task (32) subjects make choices between a fixed large amount of money at a fixed delay and an immediate amount that starts as half the delayed amount and is adjusted either up or down depending on whether the subject chooses the delayed or immediate amount in each trial, respectively. The amount of adjustments starts at half the immediate amount and is halved at each adjustment. This is repeated for five choices for each seven fixed later delays. The last choice in the procedure is used to estimate the subject's hyperbolic discount rate (or Effective Delay at 50%). Historically this task is adapted from procedures used in animals (33). We used three versions that differed in the fixed large amount: small ($10), medium ($1000), and large ($1,000,000). One random trial was chosen and contributed to the total bonus the participant received[**]. The dependent measures used for the reliability calculations were the hyperbolic discount rate for each condition as well as the area under the curve (AUC) for the small condition only.

Based on the placebo results in appendix A of Richards et al. (34) where this titration procedure was initially described for humans, retest *r*'s for discount rates for the small amount condition (n=24) was 0.939. The measurements were 2 to 5 days apart. Richard et al.'s task, however, typically involves over 100 trials (depending on how many one can complete within a

[**] Incentive compatibility in delay discounting trials is often ensured by instructing participants that a random trial will be selected and the chosen amount will be delivered at the specified delay for that option. Due to the logistical complexity of this approach,we instead added a proportional amount to each subject's bonus based on a random choice that they have made but delivered together with the rest of their compensation at the completion of the study rather than the specified delay.

given time frame), much larger than the number used in our study. In a slightly adapted version of the amount adjustment procedure, Baker, Johnson and Bickel (35) found one week *r*'s for hypothetical monetary gains ranging from 0.82 to 0.90 for non-smoking participants and from 0.71 to 0.78 for smoking participants with 30 participants in each group for the different large amount conditions. Johnson, Bickel and Baker (36) reported *r*'s of hyperbolic discount rates ranging 0.55 to 0.72 for 30 light smokers. Reed and Martens (37) reported 1 week □'s ranging from 0.68 to 0.90 for indifference amounts for the medium reward ($100) at eight delays for 46 sixth graders. Smits and colleagues (38) reported a one week *r* of 0.86 for 49 subjects using the small amount and AUC as their dependent measure. Using the original procedure of Richards et al. (34), Weafer et al. (13) reported an *r* of 0.89 for AUC of 112 subjects with no significant change between the two measurements (mean delay 8.6 days). De Wilde et al. (39) found that discount rates for 37 recovering addicts do not significantly change across four weeks and report *r*'s ranging from 0.41 to 0.50 depending on the size of the larger reward ($10, $30, $100). For 48 obese patients in a control group of a mindful eating intervention Hendrickson and Rasmussen (40) reported *r*'s for hyperbolic discount rates of 0.77 with an mean delay of 9 days using only the smallest amount ($10). Yoon and Chapman (41) reported a same day □ of 0.68 for hyperbolic discount rates using the medium amount ($100) for 220 participants but they also found a significant increase in discount rates in second measurement (the delay between the two measurements is unclear). High retest reliabilities have been reported not only with healthy control but schizophrenia patients as well: Horan, Johnson and Green (42) found a *r* of 0.67 (for AUC) for 131 patients one month apart using the large amount ($1000).

**Delay Discounting Titrator:**

In this task (43) subjects choose between a sooner smaller monetary amount and a larger later one. Unlike the other two intertemporal choice tasks in our battery, the options in this task

are more variable across participants. The sooner reward can be immediate or delayed two weeks, and the later reward can be either two or four weeks later than the sooner reward. Unlike the other delay discounting tasks in our battery, where all subjects choose between the same monetary amounts, the sooner amounts in this task are drawn from a normal distribution with a mean of 20 and standard deviation of 10, clipped at 5 and 40. The relative increment for the later versus the sooner reward can be 1, 5, 10, 15, 20, 25, 30, 50, 75% higher. Thus the options that each subject encounters differ from each other. Subjects make 36 choices. We use the percentage of larger-later (i.e. patient) choices as the dependent measure from this task.

Though other intertemporal choice tasks have been investigated for their test-retest reliabilities as detailed elsewhere in this paper, we did not find any relevant data for this particular version of the task.

**Kirby Delay Discounting Items:**

This task (44, 45) is one of the most commonly used intertemporal choice tasks that is based on the multiple price list methodology in the economics literature. Similar to other intertemporal choice tasks in the battery, subjects make choices between smaller immediate monetary amounts and larger delayed monetary amounts. The stimuli are grouped into three sets (small, medium, large) depending on the size of larger reward with nine choices in each group. Each of these nine choices span the same range of implied hyperbolic discount rates if they were to be the indifference points for a given subject (0.0016-0.025) that are spaced equidistantly on a log-scale of hyperbolic discount rates. One random trial was chosen and contributed to the total bonus the participant received (see Adaptive Adjusting Amount Delay Discounting Task for detail). We calculated both exponential and hyperbolic discount rates using both all the trials as well as for each reward size separately.

Kirby (45) reported *r*'s ranging from 0.63 to 0.77 for samples ranging from 37 to 81 people for overall hyperbolic discount rates depending on the delay between the two measurements (1 year vs. 5 weeks). Mean reliabilities for large rewards were numerically lower compared to reliabilities for small and medium rewards but this was not statistically significant. Black and Rosen (46) reported an *ICC* of 0.56 for 90 subjects for measurements that are 4 weeks apart though the details of this study are somewhat unclear. For 53 subjects re-tested between five to ten weeks apart Wölbert and Riedl (47) reported □'s of 0.61-0.68 depending on the size of the rewards. Notably these are for exponential and not hyperbolic discount rates. Yoon and Chapman (41) found same day □'s of 0.65 for discount rates of 220 participants with no significant changes across this time span using only large and small reward items. Arfer and Luhmann (48) reported *r*'s for 93 subjects of 0.85 for two measurements with a brief delay and 0.77 after a month.

**Dietary Decision-Making Task**

This task (49) consists of two phases. In the first phase subjects rate the healthiness and tastiness of fifty food items on a five point scale. A reference item that falls towards the middle of these ratings is selected. Specifically, we chose the item that was closest to the median healthiness and tastiness value of all food items. In the second phase they are given a choice between this reference item and the remaining forty nine items and rated whether they would prefer the current item over the reference item on a five point scale (Strong No, No, Neutral, Yes, Strong Yes). We use the proportion of healthy choices as well as the health and taste sensitivity as the dependent measures from this task. The latter two are operationalized as the coefficients when regressing the choice from the test phase on the health and taste differences of the presented options.

Similar to other behavioral tasks developed to study very specific questions instead of broad cognitive abilities, this task also lacks data on test-retest reliabilities in the literature. For a very similar though not identical task, Vlaev et al. (50) found an *ICC* of 0.75 and pairwise session correlations ranging from 0.42 to 0.61 for proportion of healthy choices made by 27 participants tested three times approximately a week apart between consecutive sessions. Participants had however taken different medications before two of the sessions and while a noradrenaline reuptake inhibitor did not change the behavior or the ratings, a serotonin reuptake inhibitor increased health conscious decisions as well as health ratings.

Digit Span

In this task (51) subjects view a series of digits in each trial and are asked to enter them using the mouse in the order they have seen. Subjects first complete fourteen trials reporting the digits in the order they have seen and then fourteen trials reporting the digits in the reverse order. The number of digits started at 3 and increased by 1 if the participant entered the correct series. The number of digits decreased by 1 after two incorrect responses. We use the forward and reverse digit span as the dependent measure defined as the maximum number obtained over the 14 trials.

Baddeley, Gardner and Grantham-McGregor (52) reported an *r* of 0.73 for forward span and 0.67 for backward span of 145 Jamaican children tested 3 months apart. For an auditory version of the task using only the forward condition, Karpicke and Pisoni (53) reported an *r* of 0.73 for 43 subjects tested a week apart. Williams et al. (54) reported *r*'s of 0.35 for 25 depressed patients and 0.67 for 27 healthy controls for forward span and 0.55 for depressed patients and 0.78 for normal controls for reverse span tested two weeks apart (non-computerized). Sternberg et al. (55) reported a 1-week *r* of 0.76 for forward span of 45 Tanzanian children. Though the details of the task are unclear, Gray (56) reported *r*'s of 0.85-0.87 for 22 children with normal

language abilities and 0.48-0.57 for 22 children (ages 4-6) with specific language impairments for three measurements (the second a day and the third a week later). For a sample of 663 children between the ages of 4 and 6, Alloway et al. (57) reported *r*'s of 0.81 for verbal (non-computerized) versions forward digit span and 0.53 for reverse digit span with a two week delay. For 21 subjects tested 4 weeks apart during EEG recording, Williams et al. (23) reported *r*'s of 0.50 in forward span and 0.60 in reverse span for maximum number of digits recalled without error. For 59 older adults tested approximately 17 months apart, Tröster, Woods and Morgan (58) reported □'s of 0.71 and 0.67 for forward and backward span respectively. Alloway, Gathercole and Pickering (59) found test-retest reliabilities for a verbal version of 0.84 for forward span and 0.64 for reverse span for 105 children (ages 5 to 12) tested four weeks apart. Slade et al. (60) reported *r*'s ranging from 0.27 to 0.54 for forward span of over 400 Portuguese children across a five year window with reliabilities between consecutive years increasing with time, though the details of the version of the task are unclear. They also reported significant changes in raw scores across time but these are not significant when scaled for age. For 211 first graders, Ven et al. (61) reported a test-retest correlation of 0.37 for the reverse span condition that were collected eighteen months apart.

**Directed Forgetting Task**

In this task (62, 63) subjects are presented with six letters forming two rows in each trial. After the brief presentation of the letters a cue indicates whether the top or the bottom row should be forgotten. Then a single letter is presented and subjects indicate using one of two buttons whether the letter is in their memory set (the row instructed not to be forgotten). Trials are either "positive" (the letter is in the memory set), "forget" (the letter is in the "to be forgotten" set) or "control" (the letter was not shown at all on that trial). Subjects completed three rounds of twenty four trials. The dependent measures chosen from this task were accuracy, median

response time and DDM parameters for all trials and the directed forgetting effect (forget - control trials) using the same measures.

We did not find any data on the retest reliability for this version of the task.

**Dot Pattern Expectancy Task**

In this task (64), which is an adaptation of the AX continuous performance task (65, 66), participants see cue-probe pairs that are configurations of dots on each trial. Each trial consists of the presentation of one of six cue stimuli followed by the delayed presentation of one of six probe stimuli, followed by a response. One pair consisting of a target cue (A) and a target probe (X) is considered the "target pair" (AX trial), and is identified to the participant at the beginning of the task. When the target cue is followed by the target probe the participant is asked to respond using one key and to use another key for all other cue-probe pairs (referred to as "BX" for any non-A cues followed by the probe, "BY" for any non-A cues followed by a non-X probe, or "AY" for A cues followed by non-X probes). There are 32 trials in each block and four blocks following a practice block. 68.75% of trials were AX (target) trials, 12.5% were BX, 12.5% were AY, and 6.25% were BY. The raw dependent measures from this task are the accuracy and median response time. We also calculated DDM parameters for all trials. In addition, we fit the HDDM with trial type (AX, AY, BX, BY) as a categorical predictor of drift rate. Differences in drift rate between AY and BY trials is putatively related to proactive control (AY - BY), while differences between BX and BY is putatively related to reactive control (BX - BY). We also calculated d' and bias across all trials, which are functions of participant hit rates and false-alarm rates.

For thirty eight healthy control subjects Jones, Sponheim and McDonald (67) reported test-retest reliabilities (*ICC*) of 0.71 for AX-errors, 0.39 for AY-errors, 0.28 for BX-errors, -0.06 for BY-errors, 0.66 for AX-rt, 0.63 for AY-rt, 0.34 for BX-rt, 0.54 for BY-rt and 0.80 for d’ for a

long version of the task with 40 trials per block. The reliabilities for a shorter version of the task with 32 trials per block were: 0.74 for AX-errors, 0.27 for AY-errors, 0.42 for BX-errors, -0.06 for BY-errors, 0.66 for AX-rt, 0.56 for AY-rt, 0.34 for BX-rt, 0.54 for BY-rt, 0.77 for d'. The mean interval was 23 days between the two measurements. Strauss et al. (68) reported test-retest *r*'s of 0.73-0.78 (ICC 0.7-0.77) for d' of 86 schizophrenia patients. These reliabilities did not differ from those of normal controls. Subjects completed three sessions; the first a week apart and the second two weeks apart. Their task had 104 AX, 16 BX, 16 AY and 8 BY trials. They also found improvements in performance primarily for older subjects.

**Go/no-go Task**

In this task (20), subjects emit responses to certain stimuli and omit responses to others. In our version of this task subjects see one of two colored squares. They are instructed to respond as quickly as possible by pressing a button for one color and to withhold their response for the other color. They complete ten practice trials with feedback and 350 test trials without feedback. 90% of the trials were go trials and remainder were no-go trials. Stimuli were presented for a maximum of 750 ms. The dependent measures we chose from this task are the accuracy, bias to respond, d-prime, median response times in go trials, and commission (i.e., responses on no-go trials) and omission (i.e., non-responses on go trials) errors.

Kertzman et al. (69) reported *r*'s of 0.88 for mean response times, 0.85 for the mean of the standard deviation of response times and 0.64 for commission errors of 109 participants with a version of the task that contained 20% no-go stimuli repeated a month later. Though the details of the version of their task is unclear, Weafer et al. (13) reported an *r* of 0.65 for commission errors of 123 participants tested with a mean intersession interval of 8.6 days (sd = 7.8 days). Bender et al. (70) reported an *r* of 0.62 for commission errors of 66 participants tested a week apart in a task with 25% no-go trials. Jones et al. (25) varied the go trial probability and reported

*ICC*'s of 0.836 for 20% Go RT, 0.252 for 20% Go commission errors, 0.847 for 80% Go RT and 0.548 for 80% Go commission errors of 96 subjects tested a day later. For 47 subjects in one study and 57 in another tested three weeks apart, Hedge, Powell and Sumner (71) reported *ICC*'s of 0.74 and 0.63 for go RTs, 0.76 and 0.76 for commission errors and 0.69 and 0.42 for omission errors. Their task contained 75% go stimuli which were one of three letters.

**Hierarchical rule learning Task**

In this task (72) subjects respond to eighteen different stimuli (3 shapes x 3 orientations x 2 colors) using one of three buttons. There are two rule sets. In the flat rule set each stimulus response pairing has to be learned individually based on trial-by-trial feedback. In the hierarchical rule set a hierarchical relationship between the stimuli and the correct responses allows a two-step policy where the color indicates which other dimension (e.g., the shape) should be used to determine the correct response. Subjects complete 360 trials per rule set with five breaks. The dependent measure we chose from this task was the overall accuracy. This is not the typical measure that can be extracted from this task. We chose this simplest measure because the task is an exploratory one compared to other self-regulation tasks and this measure is appropriately atheoretical reflecting our lack of specific hypotheses.

We did not find any reports of test-retest reliability for this task.

**Holt and Laury Titrator**

In this task (73) subjects choose between two gambles for ten questions: a *safe* gamble in which the two outcomes have low variance ($80 and $100) and a *risky* gamble in which the two outcomes have high variance ($190 and $5). Across the ten questions the probability of each outcome changes for both gambles. This systematic changing (i.e. titration) of the probabilities is intended to sway participants' choice from the safe to the risky gamble. The dependent measure

we chose from this task are the number of safe and risky choices as well as the risk aversion coefficient from an expected utility model.

Andersen et al. (74) reported retest reliabilities ranging from 0.34 to 0.58 for 97 Danish participants. Participants completed four versions of the titrator with amounts both equivalent to our version as well as amounts larger and smaller. These four versions were administered five times across a time span of 17 months. The correlations are those between the first session and all other sessions. There was no consistent trend of the size of the correlation depending on the size of the reward. For 53 subjects re-tested between five to ten weeks apart, Wölbert and Riedl (47) reported □ *'s* of 0.77 for curvature of value function and 0.73 for the curvature of the probability weighting function having jointly estimated the parameters. For 44 subjects Lönnqvist et al. (75) reported a test-retest reliability of 0.258 for □ and 0.205 for *r* for the number of risky decisions subjects made across around a year. Chung et al. (76) found □'s for the model-free measure of percentage of risky choices ranging from 0.36 to 0.62 for 29 to 31 healthy controls and between 0.23 and 0.51 for 47 to 65 major depressive disorder patients in four visits (mean delay 5.5 weeks). For a model-based parameter of risk aversion they found correlations ranging from 0.32 to 0.68 for healthy controls and 0.41 to 0.65 for patients.

**Information Sampling Task**

In this task (77) subjects are presented with a five by five grid of gray boxes. Each gray box can be clicked to reveal one of two underlying colors. Subjects' are instructed to indicate which color they think is in the majority. There are two conditions. In the fixed win condition subjects win or lose 100 points depending on their response regardless of how many boxes they open. In the decreasing win condition each round begins with 250 points and each opened box costs 10 points on the potential winnings of the round. An incorrect choice in this condition also leads to a loss of 100 points. Subjects complete ten rounds of each condition. The dependent

measure we chose from this task were accuracy, number of boxes opened, the mean probability of being correct in a trial given the number of opened boxes prior to the response, and the mean latency in opening each box (motivation) for both the fixed and the decreasing win conditions.

For 312 children tested at age 4, 5 and 6, Grummitt et al. (78) reported *ICC*'s of 0.55 for the fixed win condition mean response latency, 0.02 for the decreasing win condition mean response latency, 0.16 for the mean number of opened boxes in the fixed win condition, and 0.53 for the mean number of opened boxes in the decreasing win condition.

**Keep Track Task**

In this task (79) subjects are presented with a stream of fifteen words in each round where each word exclusively belongs to one of six categories (animals, colors, countries, distances, metals and relatives). Participants are instructed to remember the last word presented in a subset of those categories, which they enter in a textbox at the end of the round. Before the task begins they are given all the target categories and all possible words that might appear for each category to avoid any confusion (e.g. distances: mile, kilometer, meter, foot, inch). Each round begins by specifying which categories are relevant that round and participants complete three rounds each for three difficulty levels. The rounds differ in their difficulty based on the number of categories (ranging from 3-5). For instance, after the presentation of the number of categories (e.g. colors, relatives, animals) the presentation of the fifteen words in a trial may end with: "... dog"... "aunt"... "China"... "red"... "titanium"... "bird". At the end of the trial participants respond by typing the last word belonging to the previously mentioned targets (for this example "red, aunt, bird" as those were the last colors, relatives, and animals, respectively). The order they are written in does not matter. The score for each round is the sum of target words correctly entered into the textbox at the end. The maximum total score is therefore 36 (three repetitions of 3 points for each "3 category" round, 4 points for each "4 category" round and 5 points for each "5

category" round). The dependent measure we chose from this task was the number of correct responses.

Ven et al. (61) reported an *r* of 0.36 (though the specific dependent measure is not clear) for 211 children tested 18 months apart.

**Local-global Task**

In this task (80) participants are shown a large letter (either "H", "S", or "O") composed of smaller versions of those same letters. In each round, the color of the stimulus directed the participant to attend to either the "global" (large) letter or the "local" (small) letters. They then pressed one of two buttons to indicated whether it was an "H" or an "S" (the "O" was therefore never a response, and served as a neutral distractor). In the congruent condition the small and large letters matched, in the incongruent condition the larger letter consists of the smaller letters that would trigger the opposing response and in the neutral condition the irrelevant letter was "O", which did not trigger an alternative response. Participants completed 96 trials. The dependent measures we chose from this task were accuracy, median response time, and DDM parameters both using all trials as well as conflict (congruent, incongruent, neutral) and switch (whether global/local condition was the same or different as the last trial) contrasts. We also calculated conflict, congruent and incongruent trial response times and accuracies for the local and global conditions separately.

For a version of the task where the stimuli consisted of shapes instead of letters and the response mode was verbal, 211 children tested 18 months apart had an *r* of 0.17 in accuracy (61). Dale and Arnell (81) compared three versions of the task using different types of stimuli (faces, letters and shapes). For the letter version of the task they found *r*'s of 0.31 for the interference measures (incongruent-congruent RTs in both conditions) but 0.66 and 0.73 for RTs alone in both conditions for 55 participants returning 7-10 days later. The interference measures had

higher reliabilities for versions of the task using faces or shapes compared to versions using letter (0.70 for faces, 0.79 for shapes). These interference scores from different types of stimuli were not correlated with each other (i.e. interference scores for a version of the task using faces did not correlate with interference scores from a task using letters). A second study in the same paper found retest correlations of 0.27, 0.66 and 0.83 for the same dependent measures for another sample of 58 participants test 7-10 days apart. They find the same pattern of higher reliability for other kinds of stimuli and lack of correlation between interference scores for different versions of the tasks as well (0.57 for faces, 0.66 for shapes). For 42 subjects tested three weeks apart Hedge, Powell and Sumner (71) reported *ICC*'s of 0.69 for local congruent RT, 0.68 for local incongruent RT, 0.14 for local RT cost (incongruent RT – congruent RT), 0.56 for local congruent errors, 0.80 for local incongruent errors, 0.82 for local error cost (congruent errors – incongruent errors), 0.63 for global congruent RT, 0.70 for global incongruent RT, 0 for global RT cost, 0.60 for global congruent errors, 0.71 for global incongruent errors, 0.17 for global error cost and 0 for the global precedence effect (local congruent RT – global congruent RT). Their subjects completed 640 trials.

**Probabilistic Selection Task**

This task (82) is divided into two stages. In the first, participants learned to choose between three pairs of abstract shapes based on their reward probabilities. The reward probabilities for the shapes in each pair were 80%/20% (approach trials), 70%/30% and 60%/40% (avoid trials). Each learning block was 60 trials. Training continued for at least 3 blocks and ended when participants reached a performance criterion (greater than 70% correct on the easiest pair, 65% on the middle pair, and 50% correct on the hardest pair) or 8 blocks had passed, whichever happened first. Following this learning phase, there was a test phase where participants were shown 6 repetitions of novel pairs of stimuli that were not shown during the

learning phase (e.g. 80%/30%). The dependent measures we chose from this task are overall accuracy, accuracy for each reward probability and median response times. Two additional variables were also calculated: a general value sensitivity, and a positive learning bias. These were computed based on a logistic regression model that modeled choice (the probability of a selecting the left stimulus) during the test phase using the following formula:

$$P(left\ choice) = value\ difference * value\ sum - value\ sum + choice\ lag$$

Each stimulus value was computed based on the participant's experience with that stimulus during the training phase (rather than the objective probabilities). "Value sensitivity" was defined as the main effect of value difference. "Positive learning bias" was defined as the interaction between value difference and value sum. That is, some people may be more sensitive to value differences if both stimuli are high value, indicating that they learned the value of the "good" stimuli more effectively than the "bad" stimuli during the learning phase. The alternative is also possible - participants who learn better from negative feedback (and thus better learn the value of the low-value stimuli) would be more sensitive to value differences when the value sum is low. "Choice lag" is a nuisance variable that captures the tendency for participants to repeat their last response.

For 90 undergraduate students who completed the task twice with a delay of 7-8 weeks, Baker, Stockwell and Holroyd (83) reported r's of of 0.09 for accuracy on approach trials, -0.08 for accuracy on avoid trials, 0.269 for response times on approach trials and 0.257 for response times on avoid trials. They reported additional variables having a retest correlation <0.4 as well though it is unclear what these measures were.

**Psychological Refractory Period Task**

In this task (84–86) subjects respond to two sequential cues (a colored box is displayed, followed by a number). First they are instructed to respond using one of two buttons depending

on the color of a box. Then they are instructed to respond using one of two other buttons depending on the number that appears in the box. The interstimulus interval (ISI) between the two cues can be 50, 150, 300 or 800 ms. Subjects completed 32 trials of practice with feedback and 200 test trials without feedback. The dependent measures chosen from this task were the accuracies for each task as well as the PRP slope (the slope of regressing the task 2 response time on the ISI) and PRP slowing (difference in the task 2 response times for the 50 and 800 ms conditions).

With four blocks of forty trials and one visual and one auditory stimulus, Bender et al. (70) reported an *r* of 0.47 for the PRP effect (difference of RTs to the second cue between the short and long delay conditions) of 66 subjects tested one week apart.

**Raven's Progressive Matrices**

In this task subjects are asked to choose the item that would complete a pattern in each trial. Items increase in difficulty. The dependent measure is the number of correct responses, which is a measure of fluid intelligence and thought to reflect the ability to infer abstract rules and reason about them to solve problems.

The literature is rich with retest reliability data of this measure. Here we present a representative sample.

Raven (87) reported retest reliabilities ranging from 0.83 to 0.93 for various age groups though sample sizes and delays between measurements are unclear for these results. For twenty participants tested a week apart Watts, Baddeley and Williams (88) reported *r*'s of 0.86 and 0.91 for a paper based and computerized version of the task, respectively. Similarly Calvert and Waterfall (89) reported *r*'s ranging from 0.82 to 1 for 31 participants re-tested 4 to 8 weeks apart. For 38 participants tested for the second time two weeks later Bors and Stokes (90) found an *r* of 0.83. For 217 Guatemalan subjects re-tested about a decade later Choudhury and Gorman (91)

reported an *r* of 0.87 for accuracy. Arthur et al. (92) found a one week *r* of 0.76 for 71 participants using a shorter form test. Using a colored version of the task for 50 Kenyan children (ages 6-10) Costenbader and Ngari (93) reported two week product moment correlation of 0.84. Bors and Vigneau (94) reported *r*'s ranging from 0.85 to 0.91 for 67 participants tested on three occasions approximately 45 days apart from each other using both a short and a long form. Williams and McCord (95) found *r*'s of 0.952 for 10 subjects completing a computerized version of the task, 0.826 for 11 subjects completing a paper-based version of the task, and 0.594 for 25 subjects completing the task once with each method. Subjects were tested on average 53 days apart.

**Recent Probes Task**

This task is an updated version of the Sternberg memory scanning task (96), where subjects indicate whether a probe stimulus was part of a memory set. In our version, subjects are presented with six letters displayed in two rows. After a delay following the presentation of this memory set, subjects are presented with a single letter and asked to indicate whether the single letter was in the memory set (positive probe) or not (negative probe) using one of two buttons. Subjects complete twenty four trials per run for three runs. Half of items in each memory set were present in the previous memory set while the other half are novel. The critical update first implemented by Monsell (97) extending the task to compare probes that were included in the memory set of recent trials to probes that were not included in the memory set of any recent trials. Specifically, probes could be a member of current memory set but not of last two memory sets (positive-not-recent), a member of current memory set and of previous memory set (positive-recent), a member of previous memory set but not of current memory set (negative-recent), or a member of neither of the last two memory sets (negative-not-recent). The dependent

measures chosen from this task were accuracy and median response times and DDM parameters for all trials as well as the proactive interference contrast (negative-recent - negative-not-recent).

Data on the retest reliability for the version of the task that we used was not available in the literature. Only Barch et al. (98) report some unpublished results of quasi retest-reliabilities, which fall out of the criteria we set to include in this review. Retest reliabilities for measures from the earlier Sternberg task, however, are available. We do not present them here because the tasks are intended to capture fundamentally different cognitive processes.

**Shape Matching Task**

In this task (99) subjects indicate whether a white shape on the right of the screen and the green shape on the left of the screen are the same using one of two buttons. On half of the trials a red shape appears overlaid with the green shape. The response does not depend on this red shape. The red shape can be identical to or different from the green shape. Subjects complete forty trials for seven types of trials depending on the relationship between the target and the probe, target and the distractor and distractor and the probe. One dependent measures we calculated from this task was negative priming (the difference between median response times of trials where the target was the distractor in the previous trial versus trials where the target was not the distractor in the previous trial). Using trial-by-trial response time and accuracies, we calculated individual DDM parameters. In addition, we fit the DDM with condition (the seven relationships between the target, probe, and distractor) as a categorical predictor of drift rate. Stimulus interference was calculated as the difference in drift rate when there was a distractor present (that did not match the target or probe) and when there was no distractor present.

In the original paper describing the creation of this task, DeSchepper and Treisman (99) reported a correlation of 0.348 for the negative priming effect at the same lags between two sessions ranging from one day to one month and a correlation of 0.508 between the priming in

the first session at one lag and at lags in other sessions (n=86). We did not find any retest data reported elsewhere.

**Shift Task**

In this task (100) participants are presented with three stimuli that are each composed of one of three features from three dimensions (pattern, color, shape). The combination of features changes from trial to trial. On each trial, participants choose one of the stimuli, which results in winning 1 or 0 points. On each trial one feature is more likely to be rewarded than the other two (e.g. red), resulting in a point 75% of the time the participant chooses the relevant stimulus, compared to 25% of the time for the other two stimuli. This relevant feature stays consistent for 15-25 trials, and then switches with no external cue to the participant. Thus the participant must infer that the most rewarded feature has changed based on feedback, and relearn the important feature. Subjects complete 410 trials. The raw dependent measures chosen for this trials were the accuracy (chance being 33%) and median response time. The task was also analyzed using logistic regression and and reinforcement learning (RL) model. The logistic regression modeled the probability of a correct response using the following equation:

$$P(correct) = trial\ since\ switch * trial\ \#$$

The main effect of trials since switch was taken as a measure of learning speed, while the interaction was taken as a measure of "learning to learn".

Though no retest data are available for this specific version of the task, the task closely resembles the Wisconsin Card Sorting task which has been used for decades and has a large body of literature associated with it. Here we present a representative sample of these results using different populations. Paolo, Axelrod and Tröster (101) administered a standard version of the task to 87 older adults on two occasions separated by a little over a year. They found $\tau$'s of 0.65 for number of categories achieved, 0.16 for trials to complete first category and 0.12 for learning

to learn and *r* of 0.66 for total number of errors. In response to these results, Ingram et al. (102) tested 29 participants on average 12 days apart and reported □'s of 0.78 for total trials, 0.34 for total correct, 0.79 for total errors and 0.61 for learning to learn. Greve et al. (103) found *r*'s (corresponding *ICC*'s reported in parentheses) for 34 patients with traumatic behavioral injuries tested on average 66 weeks apart of 0.78 (0.74) for total errors. For a shorter version of the task this changed to 0.78 (0.65). For one version of the task Bird et al. (104) reported *r*'s of 0.34 for total errors achieved for 90 subjects tested about a month apart. For 54 older adults tested approximately 17 months apart Tröster, Woods and Morgan (58) reported □ of 0.37 for number of trials.

**Simon Task**

In this task (105) subjects responded using one of two arrow buttons depending on the color of the box presented right or left of center on the screen. In the congruent condition the side of the screen matched the response button, in the incongruent condition it did not. Subjects completed twenty five trials for each condition. The dependent measures chosen from this task are accuracy and median response time for all, congruent, incongruent trials and the Simon effect contrast (incongruent-congruent) as well as DDM parameters for all trials and the Simon effect contrast.

For an affective version of the task where subjects were instructed to indicate positive or negative depending on the letter type for four classes of stimuli (negative, positive, spider, general threat), De Jong et al. (106) reported *r*'s of 0.10 for the general affective Simon effect for 37 undergraduate women subjects tested a week apart. For a version of the task where the stimuli were animals instead of colored boxes Ven et al. (61) reported an *r* of 0.15 for accuracy of 208 children that were tested 18 months apart. Wöstman et al. (107) reported *ICC*'s of 0.94, 0.90, 0.68 for mean response time, standard deviation of response times, and percentage of correct

responses respectively in the congruent condition, 0.96, 0.81, 0.85 in the incongruent condition and 0.89, 0.47, 0.77 for the difference between the two conditions. Their sample consisted of twenty-three subjects who have completed the task with delays ranging from 28 to 105 days. Their task utilized triangular stimuli and subjects were instructed to respond to the direction of the triangle for 220 tasks 60 of which were incongruent. They also analyzed test-retest reliabilities for each quartile of the task and found that the retest reliabilities increased for the second half of the task. Linck and Weiss (108) reported an *r* of 0.62 for the Simon effect for 25 students tested approximately 8 weeks apart. For 75 participants tested one week apart and using letter stimuli instead of color boxes Paap and Sawi (19) found *r*'s of 0.654 for RTs in neutral baseline trials, 0.719 for incongruent trials, 0.714 for congruent trials, 0.738 for global RT and 0.428 for the incongruent congruent difference.

**Simple Reaction Time**

In this task (20) subjects are instructed to respond as quickly as possible when a stimulus ('X') is presented on the screen. They complete three blocks of fifty trials. The dependent measure for this task is the mean response time for all trials.

Kennedy et al. (22) reported an *r* of 0.59 for sixteen subjects tested for ten sessions in ten days. Choudhury and Gorman (91) found *r*'s of 0.73 of mean simple reaction times for 217 Guatemalan subjects tested almost a decade later. For 103 older adults tested 12 times in the same day Collie et al. (109) reported *r*'s ranging from 0.46 to 0.77 with lowest reliabilities between the first and second tests in each session. Erlanger et al. (110) reported a two week test *r* of 0.7 for 175 high school aged subjects. For 18 middle aged subjects tested three times two weeks apart from each other Lemay et al. (111) found an *ICC* of 0.8. Falleti et al. (112) report *ICC*'s ranging from 0.73 to 0.94 for 45 subjects tested four times within ten minutes apart and one more time a week later. For twenty participants Deary, Liewald and Nissan (24) reported *r*'s

of 0.64 for the mean and 0.47 for the standard deviation of response times. With 50 active duty military personnel sample Cole et al. (113) reported one month *ICC*'s (corresponding *r*'s in parentheses) of 0.6 (0.65) and 0.4 (0.41). Jones et al. (25) reported *ICC*'s of 0.825 for mean RT and 0.57 for RT preparation effect (faster RTs for longer response-stimulus lags) of 95 subjects tested a day later.

**Spatial Span**

In this task (114) subjects see a grid of squares in each trial. A sequence of squares is flashed red in each trial. Subjects are asked to indicate the sequence that flashed in the order presented for half of the trials and in the reverse order for the other half of the trials. They complete 14 trials per condition and receive feedback after each trial. Trials begin with a sequence of three squares, increasing in length for every correct response and decreasing for every two incorrect responses. This is a computerized version of Corsi's block tapping task. The dependent measure from this task is the forward and reverse span, computed as the maximum sequence length attained during the task.

For 1122 children aged 11 to 16 and tested 3 to 15 days apart, Orsini (115) reported *r*'s ranging from 0.7 to 0.79 depending on the age group. Baddeley, Gardner and Grantham-McGregor (52) reported an *r* of 0.6 for 145 Jamaican children tested 3 months apart. Lowe and Rabbitt (116) reported *r*'s of 0.64 for 162 subjects re-tested a month later. For 10 healthy controls re-tested 1-2 months later, Cho et al. (117) reported an *r* of 0.70. Sternberg et al. (55) reported a 1-week r of 0.41 for 45 Tanzanian children. Saggino et al. (118) reported two-week *r*'s as 0.85 for 104 older adults between 65-74 and 0.75 for 99 subjects above 74 years of age. For 21 subjects tested 4 weeks apart during EEG recording Williams et al. (23) found an *r* of 0.59 for total correctly reported sequences. For 64 children tested about a month apart Fisher et al. (119) reported an *ICC* of 0.51. In their comprehensive review, Lo et al. (120) reported

Wecshler's (121) findings of 0.72 correlation between two measurements taken 2-12 weeks apart for 141 younger subjects, of 0.70 for 156 older subjects as well as their own findings of mean correlations of 0.64 and 0.66 for cohorts of 200 middle aged adults tested three years apart.

**Stop Signal Tasks**

We included three different types of stop signal tasks in our battery:

**Classic stop signal task:**

In this task (122–124) participants are shown four different stimuli, which are each associated with one of two responses associated with the left and right hand. Participants are instructed to respond to the stimuli as quickly as possible without sacrificing accuracy. On some trials a red star appears around the stimulus as the participant prepares their response. Participants are instructed to withhold their response if they see the red star. The delay after which the stop signal appeared (stop signal delay) was adjusted using a one-up, one-down staircase procedure in 50 ms increments. This task had two conditions which differed based on how frequent stop trials were (40 % or 20 % of trials, high and low conditions). Participants completed 5 blocks of 60 trials each for each condition (the order of the two conditions was randomized across participants). The dependent measures from this task are commission and omission errors (from both go and stop trials), DDM parameters for all trials, median response time for go trials, inhibition slope (slope of the logistic regression of responses over stop signal delay), proactive slowing (comparing go RT between the two conditions, high-low), proactive SSRT speeding (comparing SSRT between the two conditions, low-high), integration SSRT (nth fastest RT - mean SSD where n is % of failed stop trials) both for all trials as well as for each condition, stop accuracy, median response time for commission errors.

For 31 children (ages 6-16) tested with a mean delay of 107 days, Kindlon, Mezzacappa and Earls (125) reported *r*'s of 0.79 for mean probability of inhibition, 0.72 for the slope of

inhibition depending on the stop signal delay, 0.61 for commission errors and 0.66 for go reaction time standard deviations. For 18 Dutch children, Kuntsi et al. (126) reported two week *ICC*'s of 0.52 for mean probability of inhibition, 0.35 for mean reaction times, 0.64 for SD of reaction times, 0.41 for total errors (only for go trials), 0.22 for commission errors, 0.37 for omission errors, 0.29 for the slope of the inhibition function and 0.11 for SSRT. Soreni et al. (127) found *ICC*'s of 0.72 for SSRT, 0.62 for go RT and 0.74 for go RT SD of 12 children with ADHD (ages 9-15) across three sessions separated by a week using an auditory stop signal and 128 trials and two go cues. Weafer et al. (13) reported an *r* of 0.65 for SSRTs with a significant decrease between the two time points (n=121; 9 days). Wöstman et al. (107) found *ICC*'s of 0.92, 0.92 and 0.29 for the mean go reaction time, standard deviation of go reaction time and the stop signal reaction time (n=23, delay range 28-105 days). While the test-retest *r*'s for each quartile of the task consistently increased for mean go and stop reaction times they remaining consistently lower than when using all trials for SSRT. For 312 children tested at age 4, 5 and 6 Grummitt et al. (78) reported *ICC*'s of 0.42 for mean reaction times and 0.62 for the proportion of successful stops. Bender et al. (70) found *r*'s of 0.60 for SSRT's of 66 participants tested one week apart. For 45 subjects in one study and 54 in another tested three weeks apart Hedge, Powell and Sumner (71) reported *ICC*'s of 0.35 and 0.57 for go RTs, 0.34 and 0.54 for mean SSD, 0.47 and 0.43 for SSRT calculated using the mean method (mean go RT - mean SSD) and 0.36 and 0.49 for SSRT calculated using integration the method.

**Motor Selective Stop Signal Task:**

In this task (128) subjects are instructed to respond using different buttons for four different stimuli as fast as possible. In a minority of trials a red star (stop signal) appears around the stimulus as the subject prepares their response. Subjects are instructed to withhold their response if they see this red star and the correct response is a designated one of the two options

(the *critical* response), but not if the red star appears but the correct response is the other option (the *noncritical* response). The delay after which the stop signal appeared (stop signal delay) was adjusted using a one-up, one-down staircase procedure in 50 ms increments. Participants completed 5 blocks of 60 trials each. 30% of the trials were "critical go" trials (no signal occurred for the critical response), 30% of trials were “non-critical go” trials (no signal occurred for the non-critical response), 20% were "critical-stop" trials (where the stop signal was shown for the critical response), and 20% were "non-critical stop" trials (where the stop signal was shown for the non-critical response). The dependent measures we chose from this task were DDM parameters, stop signal reaction time (SSRT, calculated using the integration method, Logan & Cowan, 1984, with only critical go trials acting as the underlying go distribution), accuracy and median response time on non-critical stop trials, as well as two measures we call reactive (difference between median response times of non-critical stop and non-critical go trials) and selective proactive control (difference between median response times of critical go and non-critical go trials).

We could not find test-retest reliability data on this specific version of the task. We elaborate on stop signal task reliability in a separate section below.

**Stimulus Selective Stop Signal Task:**

In this task (129, 130) participants are shown four different stimuli, which are each associated with one of two responses associated with the left and right hand. Participants are instructed to respond to the stimuli as quickly as possible without sacrificing accuracy. On some trials a red star (stop signal) or an orange star ("ignore" signal) appears around the stimulus as the participant prepares their response. Participants are instructed to withhold their response if they see the red "stop" star, but not the orange star. The delay after which the stop signal appeared (stop signal delay) was adjusted using a one-up, one-down staircase procedure in 50 ms

increments. Participants completed 5 blocks of 60 trials each. 60% of the trials were "go" trials, 20% were "stop" trials, and 20% were "ignore" trials. The dependent measures from this task are DDM parameters for all trials, SSRT using the integration method, accuracy for ignore trials, mean and standard deviation of response time for ignore trials and drift rate for the reactive control contrast (ignore trials drift rate - go trials drift rate).

We were unable to find data on the test retest reliability of this version of the task.

**Stroop Task**

In this task (131) subjects were instructed to respond using one of three keys depending on the ink color of the word presented. In the congruent condition the word matched the ink color, whereas in the incongruent condition they are mismatched. There were 96 trials (8 repetitions of each of 6 incongruent pairs and 16 of each of 3 congruent pairs, resulting in 50% congruent trials). The dependent measures for this task are accuracy, median response time and DDM parameters for both all trials and the Stroop effect contrast (incongruent-congruent) as well as median response time and number of errors for each condition separately.

Test-retest reliabilities have been assessed extensively for the Stroop task, and generally high reliability has been reported (132, 133); here we limit our review to computerized versions of the task. For 27 children (ages 6-16) tested on average 107 days apart, Kindlon, Mezzacappa and Earls (125) reported *r*'s of 0.88 for correct word responses, 0.80 for correct color responses, 0.73 for the difference between the two, 0.25 for word errors, 0.28 for color errors, 0.54 for the difference between the two and finally 0.67 for the interference score. Notably, their task included both the Stroop task (i.e. indicate the ink color) and the reverse Stroop task (i.e. read the color word) With 45 German participants Siegrist (134) conducted multiple versions of the task with different kinds of stimuli with a short delay in between. His task had 60 trials and his control condition is non-word letter sequences that are printed in different colors. Nevertheless

he found *r*'s for overall response times of 0.84 for this control condition and 0.86 for the incongruent color word condition. Using this control (unlike our congruent color word condition) he also found a moderate retest reliability for the Stroop effect (0.68). Similar results are found for the overall response time reliabilities of other types of stimuli (e.g. taboo or self-relevant words) but the Stroop effect reliabilities vary depending on what condition is used as control. For the standard color version with 110 trials and four colors Strauss et al. (135) reported *r*'s for overall response latencies of 0.71 for congruent and 0.79 for incongruent words for 28 undergraduates that were tested one week apart. The Stroop effect, however had a lower reliability (0.46). They have found similar results for different kinds of stimuli as well. For 21 subjects tested 4 weeks apart during EEG recording Williams et al. (23) reported *r*'s of 0.70 for response times and 0.60 for errors in reporting the word and 0.85 for response times and 0.60 for errors in reporting the color with no significant session effects controlling for age and gender. Bender et al. (70) found an *r* of 0.57 for the Stroop effect of 66 participants tested one week apart. For 47 subjects in one study and 56 in another tested three weeks apart Hedge, Powell and Sumner (71) reported *ICC*'s of 0.77 and 0.72 for congruent RT, 0.74 and 0.73 for neutral RT, 0.67 and 0.70 for incongruent RT, 0.60 and 0.66 for Stroop effect, 0.36 and 0.42 for congruent errors, 0.45 and 0.51 for neutral errors, 0.62 and 0.39 for incongruent errors and 0.48 and 0.44 for Stroop accuracy. Their task contained four colors.

**Task-switching Task**

In this task (136, 137) participants respond to colored numbers (1-9) based on their color (orange or blue), magnitude (greater or less than 5), and parity. On each trial a cue informs the participant of the correct rule, which is then applied to make one of two button presses. Each rule has two cues (e.g. "orange-blue" or "color"). Cue words for each rule appear above the stimulus in each trial. On each successive trial the task and cue can stay the same, the task can stay the

same and the cue can switch, or the task can switch (necessitating a cue switch). In addition, on task switch trials the task can match the task from two trials ago ("old-task", e.g. "color" -> "parity" -> "color") or differ ("new-task", e.g. "color" -> "parity" -> "magnitude"). Thus there are four trial types which were randomly sampled across trials according to the following probabilities: task-switch-old (33%), task-switch-new (33%), task-stay-cue-switch (16.5%), and task-stay-cue-stay (16.5%). The cue-target-interval (CTI) was short (100ms) for half of the trials and long (900ms) for the other half. Participants complete 60 practice trials and 440 test trials. The dependent measures for this task are the accuracy, median response time and DDM parameters for both all trials as well as for each cue-target-interval condition separately.

We have not found any data on the test-retest reliability of this version of the task.

**Tower of London**

In this task (138) subjects are presented with a starting board and a target board, each of which has three pegs and three colored balls, with the goal of making the starting board look like the target board by rearranging the colored balls making as few moves as possible. They can move only one ball at a time and are instructed to plan their entire set of moves before moving any of them. Each trial is capped at 20 seconds. Subjects complete 12 trials of increasing difficulty (the optimal number of moves varied from 2 to 5). The dependent measures from this task are the mean number of extra moves in all trials, number of trials with optimal solutions, and the mean latency before making the first move (planning time).

Lowe and Rabbitt (116) reported *r*'s of 0.60 for the number of trials with optimal solutions of 162 subjects re-tested a month later. In developing a version of the task with higher reliability Schnirman, Welsh and Retzlaff (139) reported an *r* of 0.7 for accuracy from 34 subjects tested 5 to 7 weeks apart. They used 30 items of various difficulties selected from an initial 69. Keefe et al. (140) reported 3-day *ICC*'s of 0.66 and 0.77 for 48 and 46 schizophrenia

patients for number of correct responses and 0.83 and 0.73 for 18 and 16 control subjects. Their task included 20 trials. For 19 middle aged subjects tested three times two weeks apart from each other Lemay et al. (111) found *ICC*'s of 0.3 for mean move time, 0.33 for number of optimal solutions and 0.83 for mean planning time using 16 trials. For 18 Japanese schizophrenia patients tested within the same day Kaneda et al. (141) found an *r* of 0.72 for correct responses. Dockery et al. (142) found no significant differences between the accuracies of subjects re-tested 6 (n=9) and 12 months (n=10) apart though they did not report retest reliabilities. Though they did not report correlations or *ICC*'s Bouso et al. (143) did not find any significant changes in 10 control subjects tested two hours apart using 13 trials. Köstering et al. (144) reported *r*'s (corresponding *ICC'*s in parentheses) of 0.739 (0.734) for accuracy, 0.405 (0.390) for time taken to make first move and 0.519 (0.475) for time taken for complete movements 27 subjects tested a week apart.

**Two-step Task**

In this task (145) participants make two sequential decisions between abstract shapes overlaid on different colored backgrounds. The first decision (Stage 1) between the two abstract shapes leads to one of two second "stages" (Stage 2 or Stage 3) where the participants makes a second decision between two shapes. The decision in the second phase results in either winning a coin or not. Participants' goal is to win as many coins as possible. They are told that each shape in the first stage is more likely to lead to one second stage than the other and that these probabilities remain the same across the task. They are also told that the probabilities of winning a coin from choosing either shape in the second stage changes across the task. Participants complete 50 practice trials and 200 test trials. Total points on this task contributed to the final bonus payment. Importantly, the task is structured such that each first-step decision leads to one second-stage (set of 2 shapes) frequently (70% of the time), and the other second-stage infrequently (30%). For instance, one shape in Stage 1 may lead to Stage 2 frequently and Stage

3 infrequently. This task structure is stable throughout the experiment. On the other hand, reward probabilities associated with the Stage 2 and 3 shapes adjust gradually and continuously over the experiment, to incentivize continued learning. Thus to perform optimally at the task, a participant must learn the transition probabilities at the first stage, and use them combined with trial-by-trial updates of reward probabilities to make optimal decisions. The raw dependent measure from this task is the median response time. Three additional variables were calculated based on the following logistic regression:

$$P(stay)_t = feedback_{t-1} * transition_{t-1}$$

That is, the probability of making the same choice at *t* was modeled as a function of the interaction between feedback at *t-1* and the transition (frequent or infrequent) at *t-1*. A "model-free" index was calculated as the main effect of feedback, a "model-based" index was calculated as the interaction between feedback and transition, and a "perseverance" index was the intercept of the model. We used mixed-effects logistic regression with the full interactive model fit as a random effect across participants. Individual DVs were defined based on these random effects.

We did not find any retest reliability information on this task.

**Writing Task**

In this task (developed specifically for the present study), subjects are asked to respond to the question “What happened in the last month?” for five minutes. They are asked to write for the whole time period and stay on task. The task automatically ends after five minutes. The dependent measures from this task are the probability of neutral and positive words resulting from a sentiment analysis using an openly available API through http://text-processing.com/. No previous test-retest reliability has been reported for this or similar tasks, to our knowledge. It should be noted that our analyses of this task are preliminary due lack of a clear hypothesis

relating this task to self-regulation. We included the task in the present analyses for completeness.

**Survey descriptions**

In this section we provide brief descriptions of the surveys included in our battery and summarize the literature on their test-retest reliabilities. The literature review is summarized at the survey level in Fig 1. More details on the reliabilities of specific measures can be found in the interactive figures listed under Analyses. Unlike the review of retest reliabilities for most of the behavioral tasks, where we aimed to be exhaustive, data on such psychometric properties of the surveys are more readily available. Thus this section often presents only a representative sample of results for each survey.

**Barratt Impulsiveness Scale (BIS-11)**

BIS-11 (146) is a 30 item questionnaire using a four point scale. Factor analyses reveal six first order factors that can be further grouped into three second order factors. The second order factors are *attentional, motor* and *non-planning.* The attentional second order factor consists of the *attention* ('I "squirm" at plays or lectures') and *cognitive stability* ('I often have extraneous thoughts when thinking') first order factors. The motor second order factor consists of *motor* ('I act "on impulse"') and *perseverance* ('I change residences') first order factors. The nonplanning second order factor consists of the *self-control* ('I am a careful thinker') and *cognitive complexity* ('I like to think about complex problems') first order factors.

In their fifty year review Stanford et al. (147) reported one month □'s for 153 participants of 0.83 for total scores, 0.61, 0.67, 0.72 for the attentional, motor and non-planning second order factors and 0.74, 0.67, 0.73, 0.50, 0.35, 0.23 for the attention, motor, self-control, cognitive complexity, perseverance and cognitive instability first order factors. For an Italian version Fossati et al. (148) reported two month *r*'s ranging from 0.62 to 0.82 on the six first order factors

and ranging from 0.82 to 0.88 for the second order factors on 83 subjects (though there were some differences in factor loadings). For the total score the correlation was 0.89. Comparing different testing modes Suris et al. (149) found 2 week *r*'s (corresponding *ICC*'s in parentheses) of 0.79 (0.63), 0.66 (0.43), 0.85 (0.72), 0.88 (0.77) for the attention, motor and non-planning factors as well as the total for 32 subjects who complete only paper-based versions of the survey. These numbers changed to 0.79 (0.62), 0.50 (0.25), 0.86 (0.73), 0.88 (0.78) for a group (n=31) that completed first a paper-based then a computerized version and to 0.76 (0.58), 0.75 (0.57), 0.88 (0.77), 0.89 (0.80) for a group (n=34) that only completed the computerized version. For a Turkish version the two month reliability (n=44) of the total score was 0.83 and it ranged from 0.65 to 0.80 for the second order factors (150). Weafer et al. (13) reported (n=127; 9 days) an *r* of 0.92 for the total score, 0.86 for the attention factor, 0.88 for the motor factor and 0.85 for the non-planning factor.

**Behavioral Inhibition and Approach (BIS/BAS)**

Developed by Carver and White (1994) to measure two theoretical constructs of behavioral approach system (BAS) and behavioral inhibition system (BIS), this survey is a 24 item survey that has a four factor solution: 4 items for *BAS drive* ('I go out of my way to get things I want.'), 4 items for *BAS fun seeking* ('I'm always willing to try something new if I think it will be fun.'), 5 items for *BAS reward responsiveness* ('When I'm doing well at something I love to keep at it.') and 7 items for *BIS* ('Even if something bad is about to happen to me, I rarely experience fear or nervousness') and the remaining questions are fillers. Questions are presented with four point scales.

In the original paper Carver and White (1994) reported *r*'s of 0.66 for BIS, 0.66 for drive, 0.59 for reward responsiveness and 0.69 for fun seeking for 113 subjects tested 8 weeks apart. Sutton and Davidson (1997) reported 5 month *ICC*'s of 0.72 for BAS and 0.68 for BIS for 46

subjects. Meyer, Johnson and Winters (2001) reported one year *r*'s of 0.81 for BIS, 0.50 for BAS total, 0.44 for reward responsiveness, 0.46 for drive and 0.49 for fun seeking for 42 bipolar subjects. For four measurements across a year that were 3 months apart from each other Li and Zinbarg (2007) reported *r*'s ranging from 0.70 to 0.75 using BIS only for 109 students. Alloy et al. (2012) reported *r*'s over an mean of 1.8 months of .81 and .82, and over an average of 8.8 months of .70 and .60, for BIS and BAS total, respectively for 201 subjects in the bipolar spectrum interviewed every four months for 4.5 years. Amiri, Behnezhad and Azad-Marzabadi (2017) reported 4 week *r*'s of 0.64 for BIS and 0.58 for BAS for 70 participants.

**Brief Self-Control Scale (BSCS)**

BCSC is a 13 item survey presented with 5 point scales. It is initially developed by Tangney, Baumeister and Boone (151). Maloney, Grawitch and Barber (152) show a two factor structure representing *restraint* ('I wish I had more self-discipline') and *impulsivity* ('Sometimes I can't stop myself from doing something, even if I know it is wrong').

Tangney et al. (151) reported an *r* of 0.87 for 233 participants tested approximately three weeks later for all items. Duckworth and Seligman (153) reported 7 month *r*'s of 0.75 for 140 8th graders and 0.76 for 164 students for all items.

**Dickman's Functional and Dysfunctional Impulsivity**

This survey (154) uniquely distinguishes between two types of tendencies to act without forethought: one that has negative consequences and one that is more optimal. It consists of 11 true/false items for the *functional impulsivity* factor (e.g. 'I don't like to do things quickly, even when I am doing something that is not very difficult.' or 'I don't like to make decisions quickly, even simple decisions, such as choosing what to wear, or what to have for dinner') and 12 for the *dysfunctional impulsivity* factor (e.g. 'Often, I don't spend enough time thinking over a situation before I act.' or 'I often say and do things without considering the consequences').

Brunas-Wagstaff et al. (155) reported *r*'s of 0.58 for the functional subscale and 0.73 for the dysfunctional subscale for 211 children re-tested 3-4 weeks later using a slightly modified version suitable for the subject age range. When broken down by age group these ranged from 0.48 to 0.62 for the functional and from 0.62 to 0.73 for the dysfunctional subscales with increasing stability with age. For 49 students Caci et al. (156) reported a one year retest reliability separately for males and females. For 11 males they reported □'s of 0.435 and 0.824 for the functional and dysfunctional subscales while for 38 females these numbers were 0.696 and 0.528. For 107 students Chico et al. (157) reported one month *r*'s of 0.757 for the functional and 0.765 for the dysfunctional subscale using a Spanish translation of the survey.

**Domain Specific Risk Taking (DOSPERT)**

DOSPERT (Domain Specific Risk Taking) survey (158, 159) attempts to capture a more comprehensive, interpretable and translatable construct of risk attitude. This stands in contrast to measures inspired by the expected utility theory that reduces risk attitude to a single number across domains and does not distinguish between marginal value for outcomes and attitudes towards risk. The abbreviated version (159) consists of 30 scenarios that are presented with slight variations in question wording to form three separate subscales intended to disentangle these. In the *risk taking* subscale subjects are asked the likelihood they would engage in the described activity; in the *risk perception* subscale they are asked how risky they assess each situation to be and finally in the *expected benefits* subscale they are asked the benefit they would expect from each situation. These scenarios are chosen from five domains based on prior literature: *Financial* (F; 'Betting a day's income at the horse races.' This consists of two factors: Investing and gambling), *health/safety* (HS; 'Drinking heavily at a social function.'), *recreational* (R; 'Going camping in the wilderness.'), *ethical* (E; 'Taking some questionable

deductions on your income tax return.'), and *social* (S; 'Admitting that your tastes are different from those of a friend.'). All items are presented with a 7 point scale.

For an initial longer version Weber, Blais and Betz (158) reported one month *r*'s for 121 subjects for both the risk taking and risk perception surveys. For the risk taking survey they reported correlations of 0.44, 0.58, 0.75, 0.72, 0.80 for F, S, HS, E, R. For the risk perception survey they found correlations of 0.42, 0.47, 0.62, 0.67 and 0.56 for the same domains. For a Chinese adaptation of a longer version of the risk taking survey Du, Li and Du (160) reported 4 week *r*'s for 155 subjects of 0.81 for R, 0.60 for E, 0.72 for I, 0.49 for G, 0.70 for S, 0.70 for HS and 0.76 for the total score.

**Emotion Regulation Questionnaire (ERQ)**

Developed by Gross and John (161) the ERQ is a ten item survey that measures two emotion regulation strategies: *reappraisal* ('I control my emotions by changing the way I think about the situation I'm in') and *suppression* ('I control my emotions by not expressing them'). Items are presented on a seven point scale.

The original paper reported 3 month test retest reliabilities of 0.69 for both factors (n=791). Balzarotti, John and Gross (162) reported 2 month *r*'s of 0.67 for reappraisal and 0.71 for suppression for 182 Italian participants. For 692 Australian children (ages 10-18) Gullone and Taffe (163) found 1 year *ICC*'s ranging from 0.37 to 0.47 for reappraisal and 0.40 to 0.63 for suppression with an increasing trend with age. For a Spanish version Cabello et al. (164) reported 3 month *r*'s of 0.66 for suppression and 0.64 for reappraisal (n = 115). For a Turkish adaptation Eldeleklioğlu and Eroğlu (165) reported a 3 week *r* of 0.74 for reappraisal and 0.72 for suppression of 90 subjects. For a Persian version Hasani (166) reported 5 week *r*'s of 0.68 for suppression and 0.77 for reappraisal (n=150).

**Five Facet Mindfulness Questionnaire (FFMQ)**

FFMQ is a result of a broad psychometric analysis of multiple mindfulness questionnaires. Baer et al. (167) chose the 39 items that best loaded on the five factor solution. The five facets resulting from factor analyses are *observing* ('When I'm walking, I deliberately notice the sensations of my body moving.'), *describing* ('I'm good at finding the words to describe my feelings.'), *acting with awareness* ('I find it difficult to stay focused on what's happening in the present.'), *non-judging of inner experience* ('I criticize myself for having irrational or inappropriate emotions.') and *nonreactivity to inner experience* ('I perceive my feelings and emotions without having to react to them.'). Items are presented with a five point scale.

Using a Dutch version of the task Isenberg (168) reported two week *ICC*'s of 0.798 for the total FFMQ, 0.863 for observing, 0.820 for describing, 0.657 for acting with awareness, 0.757 for non-judging of inner experiences and 0.776 for nonreactivity to inner experiences for 30 mostly female fibromyalgia patients. Using a Chinese version Deng, Rodriguez and Xia (169) reported one month *r*'s of 0.741 for observing, 0.699 for describing, 0.436 for acting with awareness, 0.611 for non-judging and 0.512 for nonreactivity for 81 students. For 41 subjects Petrocchi and Ottaviani (170) reported 20 month *r*'s between two measurements of 0.19 for observing, 0.66 for describing, 0.79 for acting with awareness, 0.58 for non-judging and 0.59 for nonreactivity.

**Future Time Perspective (FTP)**

Developed by Lang and Carstensen (171) in the context of socioemotional selectivity theory, FTP aims to quantify the age-related changes in how people view their future in selecting their goals. It consists of 10 items presented on a five point scale. Based on their scores people are categorized into having either more open-ended or more limited time perspectives. Older

people tend to have the latter. Example items include ‘Many opportunities await me in the future’ and ‘Most of my life (still) lies ahead of me.’

Zacher and Lange (172) used three items each for two theoretical constructs of interest taken from the FTP: focus on opportunities and focus on limitations. For a sample of 85 Dutch employees they reported 3 month *r* of 0.62 for the former and 0.5 for the latter using this restricted set of the FTP items. Kooji, Bal and Kanfer (173) reported a one year *r* of 0.76 for 765 Dutch employees. We did not find any additional data using the whole set of items reporting retest reliability.

**Grit Scale (GRIT-S)**

Developed by Duckworth and Quinn (174) the short Grit scale aims to measure perseverance. It consists of eight items presented on a five point scale. Grit-S yields a two factor structure: *consistency of interest* (‘I often set a goal but later choose to pursue a different one’) and *perseverance of effort* (‘I finish whatever I begin’).

In the original paper Duckworth and Quinn (174) reported a one year *r* of 0.68 for 279 middle and high schoolers. For 121 college students Hill, Burrow and Bronk (175) reported 3 month *r*’s between all items of 0.61. For a Chinese version Li et al. (176) reported a four week *r* of 0.78 for all items 0.63 for consistency of interest and 0.70 for perseverance of effort for 138 10th graders.

**I-7 impulsiveness and venturesomeness questionnaire**

The culmination of Eysenck’s work in developing an impulsivity questionnaire (177), the I-7 is the most recent version following I-5 and I-6. Though the survey is conceived to have three components we only used the 19 items for the *impulsiveness* (e.g. ‘Are you an impulsive person’) and 16 items for the *venturesomeness* (e.g. ‘Would you enjoy the sensation of skiing very fast down a high mountain slope?’) factors omitting the empathy factor.

For 132 participants Luengo et al. (178) reported one month *r*'s of 0.76 for the impulsiveness factor and 0.80 for the venturesomeness factor using a Spanish version of the survey. We did not find any additional reliability data on this questionnaire.

**Mindful Attention and Awareness Scale (MAAS)**

Developed by Brown and Ryan (179) MAAS is a 15 item questionnaire presented on a six point scale. MAAS focuses on the 'individual differences in the frequency of mindful states over time.' These items load onto a single factor. Sample items include 'I could be experiencing some emotion and not be conscious of it until some time later.' and 'It seems I am "running on automatic" without much awareness of what I'm doing.'

Brown and Ryan (179) reported a 4 week *ICC* of 0.81 for 61 college students. Barnes et al. (180) reported 10 week r of 0.73 for 82 college students. For a Chinese version Deng et al. (181) reported 20 day *r* of 0.54 for 70 students and no significant differences between the two time points. Murphy et al. (182) reported an 8 week *r* of 0.66 for 441 women. For 46 pregnant women who completed the questionnaire 4 times a week for 3 weeks Matvienko-Sikar and Dockray (183) reported an *r* of 0.94.

**Multidimensional Personality Questionnaire (MPQ) Control Scale**

The MPQ is a comprehensive questionnaire consisting of multiple subscales. We only used items from the 24-item single factor *control* subscale, adopting the strategy of Whiteside and Lynam (184). Typical true/false items for the MPQ are 'I am fast and careless.' or 'I do things on the spur of the moment.'

According to Whiteside and Lynam (184) this survey had a one month retest reliability of 0.82 (n not reported). We did not find any additional reliability data on this survey.

**Selection-Optimization-Compensation (SOC) questionnaire**

This questionnaire is developed as a measurement tool of life management strategies within lifespan psychology (185). It is intended to measure three components: *Selection*,

*optimization* ('I keep working on what I have planned until I succeed' vs 'When I do not succeed right away at what I want to do, I don't try other possibilities for very long') and *compensation* ('When things don't go as well as they used to, I keep trying other ways until I can achieve the same result I used to' vs 'When things don't go as well as they used to, I accept it'). The selection component consists of two first level factors: *Elective selection* ('I concentrate all my energy on a few things' vs 'I divide my energy among many things') and *loss based selection* ('When things don't go as well as before, I choose one or two important goals' vs 'When things don't go as well as before, I still try to keep all my goals'). Each item presents two scenarios that the subject chooses between. There are twelve items for each component.

Freund and Baltes (186) reported one month *r*'s of 0.77, 0.71 and 0.76 for elective selection, optimization and compensation (loss based selection had not been developed at this point) for 218 German subjects. No additional retest reliability data were found for this questionnaire.

**Sensation Seeking Scale (SSS)**

Initially developed by Zuckerman et al. (187) this survey is intended to measure the concept of optimal stimulation level. Subjects are presented with two scenarios in each question and asked to indicate which they would prefer. Zuckerman (188) identified four factors that the survey measured: *boredom susceptibility* (BS; 'There are some movies I enjoy seeing a second or even a third time' vs. 'I can't stand watching a movie that I've seen before'), *disinhibition* (D; 'I like "wild" uninhibited parties' vs 'I prefer quiet parties with good conversation'), *experience seeking* (ES; 'I dislike all body odors' vs. 'I like some for the earthly body smells'), and *thrill and adventure seeking* (TAS: 'I often wish I could be a mountain climber' vs 'I can't understand people who risk their necks climbing mountains'). We used the 40 item form V with ten items for each factor.

Zuckerman et al. (189) reported one week *r*'s of 0.89 for total, 0.94 for TAS, 0.92 for ES, 0.91 for D, and 0.82 for BS for 38 subjects. Zaleski (190) reported *r*'s over a period of 6 weeks for 31 subjects of 0.83, 0.78, 0.91, 0.66 for TAS, ES, D, BS. Thombs et al. (191) reported two week *r*'s of 0.69, 0.63, 0.81, 0.66, 0.82 for TAS, ES, D, BS and total score for 61 drinkers. For an Estonian version Parmak, Mylle and Euwema (192) reported 3 month *r*'s correlations of 0.67, 0.71, 0.69, 0.57, 0.52 for total score, TAS, ES, D and BS for 87 conscripts.

**Short Self Regulation Questionnaire (SSRQ)**

The 31 item short self regulation questionnaire is an abbreviated version of the self regulation questionnaire developed by Brown, Miller and Lawendowski (193) based on Carey, Neal and Collins' (194) work. Neal and Carey (195) show that these items load onto two factors: *impulse control* ('I learn from my mistakes.') and *goal setting* ('I set goals for myself and keep track of my progress.'). Items are presented on a 5 point scale.

We could not find test-retest reliability data for this shortened version of the questionnaire.

**Stanford Leisure-Time Activity Categorical Item (L-Cat)**

The L-Cat (196) is a single item that is intended to measure physical activity level. It provides six descriptions ranging from 'I did not do much physical activity. I mostly did things like watching television, reading, playing cards, or playing computer games. Only occasionally, no more than once or twice a month, did I do anything more active such as going for a walk or playing tennis.' to 'Almost daily, that is five or more times a week, I did vigorous activities such as running or riding hard on a bike for 30 minutes or more each time.'

Kiernan et al. (196) reported a □ of 0.8 for 267 female obese patients tested 2-6 weeks apart. Riebl et al. (197) reported an *r* of 0.86 on two measurement of the L-Cat using an

interactive version on 60 participants. The measurements were separated by 3-14 days. No additional data were found for this questionnaire.

**Ten-Item Personality Inventory (TIPI)**

Developed by Gosling, Rentfrow and Swann (198) TIPI measures the Big Five personality traits of *extraversion* (E; 'Extraverted, enthusiastic'), *openness* (O; 'Open to new experiences, complex'), *conscientiousness* (C; 'Dependable, self-disciplined'), *agreeableness* (A; 'Sympathetic, warm'), and *emotional stability* (ES; 'Calm, emotionally stable'). Subjects rate themselves on combinations of two adjectives in each question using a seven point scale.

The original paper reported a six week *r*'s on 180 subjects of 0.77 for E, 0.62 for O, 0.76 for C, 0.71 for A and 0.70 for ES and. For a Spanish version Romero et al. (199) reported 6 week *r*'s of 0.79 for E, 0.78 for O, 0.69 for C, 0.52 for A and 0.83 for ES with 198 subjects. Renau et al. (200) reported one month *r*'s of 0.81, 0.72, 0.77, 0.61, 0.76, for 31 Spanish participants and 0.85, 0.70, 0.81, 0.69, 0.82, for 49 Catalan participants for the E, O, C, A, ES factors respectively.

**Theories of Willpower Scale**

Developed by Job, Dweck and Walton (201) the Theories of Willpower Scale measures beliefs about willpower and the role of ego depletion in self control. It consists of 12 items presented with a six point scale. Higher scores indicate stronger beliefs viewing self control as a limited resource. Half of the items are about *strenuous mental activity* ('Strenuous mental activity exhausts your resources, which you need to refuel afterwards (e.g. through taking breaks, doing nothing, watching television, eating snacks).') and the other half about *resisting temptations* ('Resisting temptations makes you feel more vulnerable to the next temptations that come along.').

Job, Dweck and Walton (201) reported *r*'s for 41 college students across three measurements during the academic year that are >0.77 (exact values not reported in reference). No additional data were found for this questionnaire.

**Three Factor Eating Questionnaire (TFEQ-R18)**

TFEQ-R18 is a shortened measure by Karlsson et al. (202) capturing eating behavior in both patient and healthy populations. It measures three aspects of eating behavior: *cognitive restraint* ('I deliberately take small helpings as a means of controlling my weight.'), *uncontrolled eating* ('When I smell a sizzling steak or juicy piece of meat, I find it very difficult to keep from eating, even if I have just finished a meal.') and *emotional eating* ('When I feel anxious, I find myself eating.'). 18 questions are presented on four point scales though the options for the scale ratings differ across questions.

We searched only for data on this shortened version. Mostafavi et al. (203) reported a two week *r* of 0.87 for 126 women. We did not find any additional data on this questionnaire.

**UPPS-P**

Whiteside and Lynam (184) initially developed the four factor UPPS after administering a wide variety of impulsivity surveys and combining items from each survey that loaded highest to the four factor solution. This was expanded on by Lynam et al. (204) to measure a fifth construct as well. The five factors that constitute the abbreviated name of the questionnaire are 12-item (negative) *urgency* ('I have trouble controlling my impulses'), 11-item (lack of) *premeditation* ('I have a reserved and cautious attitude toward life'), 10-item (lack of) *perseverance* ('I generally like to see things through to the end'), 12-item *sensation seeking* ('I generally seek new and exciting experiences and sensations') and 14-item *positive urgency* ('When I am very happy, I can't seem to stop myself from doing things that can have bad consequences'). All items are presented with a four point scale.

For the initial version of the survey lacking the positive urgency subscale Anestis, Selby and Joiner (205) reported 3-4 week *r*'s of 0.73 for negative urgency, 0.73 for lack of premeditation, 0.86 for sensation seeking and 0.64 for lack of perseverance for 65 students. Weafer et al. (13) reported (n=126; 9 days) *r*'s of 0.86 for the negative urgency factor, 0.81 for the lack of premeditation, 0.83 for the lack of perseverance, 0.93 for sensation seeking and 0.85 for positive urgency. They also found a significant decrease for positive urgency and significant increase for negative urgency between the two time points. For 407 students tested three times annually Kaiser (206) reported *r*'s ranging from 0.70 to 0.74 for negative urgency, from 0.81 to 0.88 for sensation seeking and from 0.66 to 0.74 for lack of premeditation (the other factors are not reported). For 50 subjects in one study and 62 in another tested three weeks apart Hedge, Powell and Sumner (71) reported *ICC*'s of 0.72 and 0.73 for negative urgency, 0.70 and 0.85 for premeditation, 0.73 and 0.78 for perseverance, 0.87 and 0.89 for sensation seeking, and 0.80 and 0.82 for positive urgency. In a third study with 42 subjects the *ICC*'s in the same factor order were: 0.78, 0.88, 0.90, 0.91 and 0.85.

**Zimbardo Time Perspective Inventory (ZTPI)**

ZTPI (207) aims to measure how people view time and how this may affect their lives in a broader context. It consists of 56 items and uses a 5 point scale. CFAs show a five factor solution for the survey: *Past-negative* (PN; 'I think about the bad things that have happened to me in the past'), *present-hedonistic* (PH; 'Taking risks keeps my life from becoming boring'), *future* (F; 'It upsets me to be late for appointments'), *past-positive* (PP; 'It gives me pleasure to think about the past') and *present-fatalistic* (PF; 'My life path is controlled by forces I cannot influence').

Zimbardo and Boyd (207) reported 1 month *r*' for 58 students of 0.80 for the F subscale, 0.76 for PF, 0.76 for PP, 0.72 for PH and 0.70 for PN. For 278 Latvian and 407 Russian high

school students four week *r*'s were 0.89 and 0.90 for PN, 0.84 and 0.73 for PH, 0.82 and 0.78 for F, 0.74 and 0.54 for PP and 0.69 and 0.81 for PF subscales of each sample respectively (208). For an extended Swedish version of the questionnaire Carelli, Wiberg and Wiberg (209) reported *r*'s of 0.85 for PN, 0.74 for PH, 0.71 for PF, 0.69 for PP, and 0.64 for F for 30 participants tested 2 weeks apart. Four week *r*'s of 76 participants were 0.73 for PP, 0.85 for PN, 0.76 for PH, 0.66 for PF and 0.77 for F (210). For 51 students Wang et al. (211) reported six month *r*'s of 0.61 for PP, 0.76 for PN, 0.68 for PH, 0.54 for PF and 0.55 for F factors using a Chinese version of the survey.

**Effect of task length on stability**

For this example we look at the retest reliability of the dependent measures from the Shift task with 410 trials. Figure 6 depicts the three patterns found for the measures in this task on how the point estimates of the retest reliability changed as a function of the number of trials used to estimate them. The average response time is the only dependent measure that reaches an acceptable, albeit moderate, level of reliability that would justify its use as an individual difference measure, It reaches its maximum reliability in half the number of trials used in this task. While most dependent measures show an increasing trend in their reliability with increasing number of trials, only two other measures (non-perseverative errors and conceptual responses) approach somewhat acceptable levels when all trials are used. The rest of the measures have reliabilities less than 0.5 regardless of the number of trials used to estimate them. Based on this analysis, a researcher might question whether to use the Shift task for individual difference analyses, as many of the dependent measures that are usually of primary interest, e.g. perseverative responses, exhibit little or no reliability even after hundreds of trials. Alternatively,

if one is interested in using one of the relatively reliable measures on this task, this analysis shows that reliable results can be obtained with relatively few trials.

**Data quality checks**

To capture real-world behaviors and provide a data quality anchor for this study, we had 74 demographic items covering a wide range of topics including alcohol use (10 questions), caffeine intake (4 questions), drug use (19 questions), finances (10 questions), mental health (7 questions), physical health (7 questions), risk taking behavior (4 questions), smoking behavior (6 questions), and social interactions (7 questions). A full list of demographics items can be found in Table S2.

We anticipated that demographic variables would show little to no change over the observed time period, so we examined their retest reliability as a post-hoc data quality check. The median reliability for all items was 0.83 (range -0.08 to 1). Primary demographics all had very high retest reliability (Ethnicity = 1, Age = 0.99, Sex = 0.99, Household income = 0.99, Highest education = 0.95, Weight = 0.95), which suggests that our participants were not responding haphazardly. Items such as frequency of hazardous cannabis usage (0.05), amount of caffeine per day from sources other than coffee (0), and spousal or parental complaints on drug use (-0.033) on the other hand had the lowest reliability, primarily due to lack of variance in response to these questions as most participants responded with 0 or 'not applicable.'

As a second check of data quality we examined the effect of the variable retest delay on the change of each subjects' scores; due to our data collection strategy (see Methods) this delay was not strictly controlled. Since we only had two measurements we could not directly test whether a measure becomes less reliable depending on the delay between the two time points; rather, we tested whether the difference score distribution for each measure depended on the

delay between the two measurements. We regressed the standardized difference scores against the retest delay allowing for random intercepts for each subject and random slopes for each measure. We found that although there were large random effects, the fixed effect of retest delay on the difference scores was negative (posterior mean for fixed effect of retest delay on standardized difference scores = -0.01, 95%, credible intervals = [-0.015, -0.005]; Fig S1). The difference score decreased slightly for subjects with longer delays between the two measurements alleviating the concern that longer delays may have led to larger differences. None of the differences scores for individual measures showed significant dependence on retest delay after Bonferroni correction.

As a third data quality check, we computed correlations between similar survey items, computing textual similarity using Levenshtein distance*[††]. We found 19 item pairs that were similar across all the surveys, using a threshold of similarity > 0.8. For example, two similar questions were a question on the BIS BAS survey ( “I often act on the spur of the moment”) and one in the BIS-11 survey (“I act on the spur of the moment”; similarity for these items was .842). The median absolute polychoric correlations between such items comparing both items for each time point ranged from 0.58 to 0.61. The correlations were marginally higher for more similar items (posterior mean for the main effect of standardized Levenshtein distance on polychoric correlations = 0.093, 95% credible interval = [0.0009, 0.184], Figure S2). For example, the polychoric correlation for the items listed above (with similarity 0.842) were 0.61 and 0.69 while for another pair of identical items (similarity = 1) they were 0.82 and 0.61 for each time point.

---

[††]* Similar survey items were defined as those that have >0.8 of a similarity metric defined as 1-d(str1, str2)/max(A, B), where d represents the Levenshtein distance between the two strings str1 and str2 and A and B are the number of characters in each string.

Thus we concluded that our participants were giving similar but not identical answers to similar questions that they encountered at different time points. These analyses provide some degree of assurance that the participants were real people and not automated machines.

**Analyses of reliability estimate variability in the literature**

While the relationship between the empirical and literature-derived reliability estimates seems weak, this must be contextualized by evaluating the variability of retest reliability estimates in the literature. If individual studies in the literature have similarly weak relationships to the literature-wide retest reliability for a given measure (i.e. if the variance of the reliability estimates reported in the literature for a given measure is large), this suggests a general issue of variability in retest reliability estimates across samples and not a specific issue with our sample. Therefore, we compared two types of models: (1) One where we predicted the literature retest reliability using an estimate sampled from the literature review (similar to leave one out cross-validation but using the left out value as the predictor instead). This reflects a noise ceiling for the literature and captures how well the literature could possibly be estimated. (2) Another model where we predicted the literature retest reliability using the estimate from our sample. In both models we also accounted for the effect of sample size and whether the measure was a task or survey measure as these were found to account for significant portions of variance. We computed the variance explained (adjusted $R^2$) for both models across 1000 iterations.

**Differences between different DDM approaches**

Comparing the different modeling approaches with each other for non-contrast measures, HDDM estimates are significantly less reliable than EZ estimates (mean of posterior for fixed effect of indicator variable for HDDM estimates = -0.107, 95% credible interval = [-0.184, -0.41]) which are comparable in reliability to raw measures (mean of posterior for fixed effect of

indicator variable for raw variables = -0.041, 95% credible interval = [-0.105, 0.020]). For contrast measures, on the other hand, the HDDM estimates are more reliable than EZ estimates (mean of posterior for fixed effect of indicator variable for HDDM estimates = 0.141, 95% credible interval = [0.037, 0.252]) and raw variables fall between the two modeling approaches (mean of posterior for fixed effect of indicator variable for raw variables = 0.085, 95% credible interval = [-0.008, 0.171]).

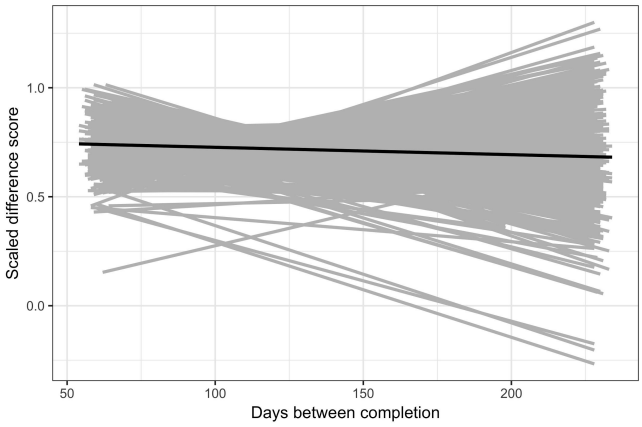



**Fig. S1.** Effect of days between completing the two measurements on the difference between the scores from the two time points. The black reflects the trend for all measures with a significant slight negative slope. None of the individual slopes is significant accounting for multiple comparisons.

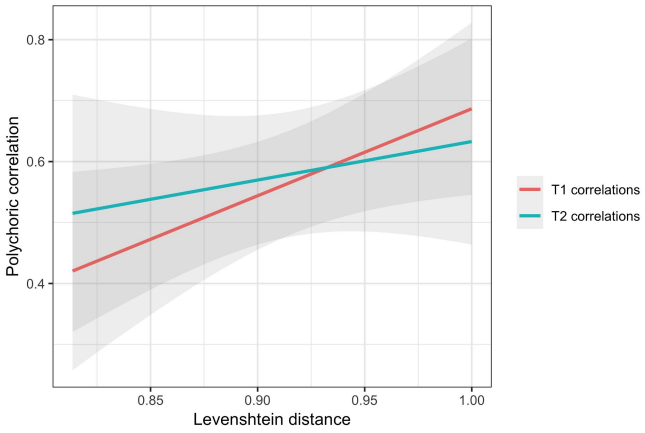



**Fig. S2.** Relationship between polychoric correlations and similarity of survey items. Red line denotes correlations between similar items in the first time point and the blue line in the second time point.

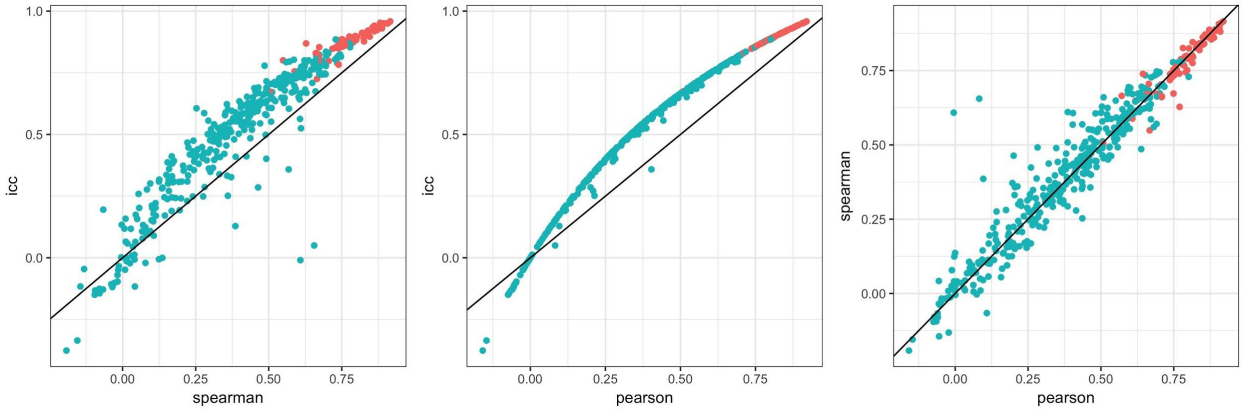



**Fig. S3.** Scatter plots of different reliability metrics compared to each other. Point estimates of ρ's, r's and ICC's for each variable are depicted. Red dots are dependent measures from survey measures while blue dots are those from behavioral tasks.

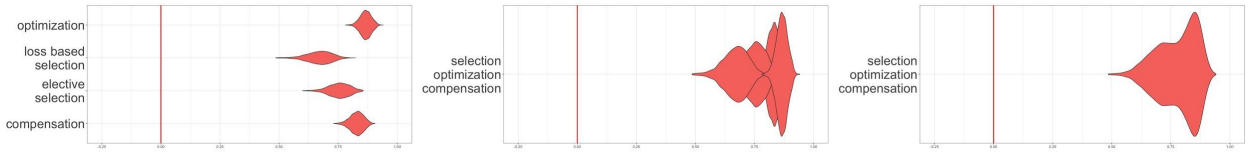



**Fig. S4.** Example of how the bootstrapped reliability distributions for multiple measures of a task are overlaid and combined when creating the violin plots for each task in Figure 4.

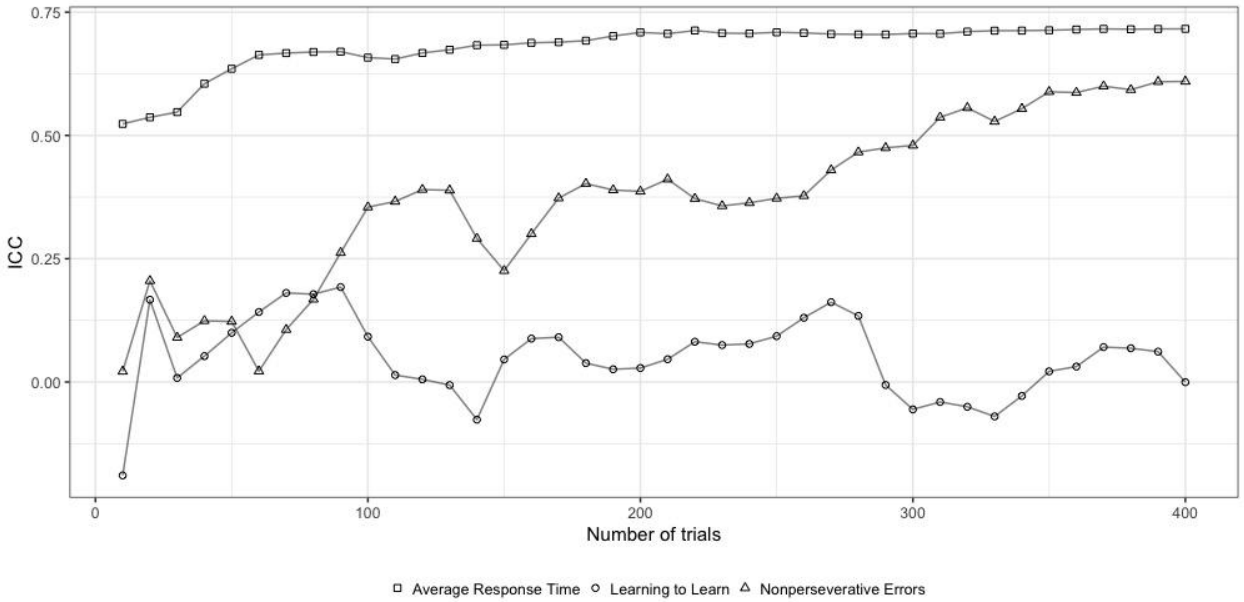



**Fig. S5.** Change in point estimates of ICC for three dependent measures selected from the Shift task using increasing numbers of trials to estimate them. These measures show the different types of relationships reliability estimates have based on the trial numbers used to estimate them.

**Table S1.** Test-retest reliability metrics

| Measure | Type | Formula | Pros | Cons |
|---|---|---|---|---|
| Pearson's $r$ | Relative | $\frac{Covariance\ of\ two\ measurements}{Product\ of\ the\ standard\ deviations}$ | - Ease of calculation and interpretation<br>- Scale independent | - Inability to detect systematic error<br>- Limited to two time points<br>- Not a measure agreement (accuracy)<br>- Sample size dependent |
| Intraclass correlations ($ICC$) | Relative | $\frac{Between\ subject\ variability}{Between\ subject\ variability + Error}$ | - Scale independent<br>- Multiple types for various scenarios | - Researcher degrees of freedom in choosing the appropriate one<br>- Not a measure agreement (accuracy) |
| Spearman's $\rho$ | Relative | $\frac{Covariance\ of\ two\ measurements'\ rank\ ordering}{Product\ of\ the\ standard\ deviations\ of\ rank\ orderings}$ | - Ease of calculation and interpretation<br>- Scale independent | - Inability to detect systematic error<br>- Limited to two time points<br>- Not a measure agreement (accuracy) |
| Kendall's $\tau$ | | $\frac{Number\ of\ concordant\ pairs - Number\ of\ discordant\ pairs}{n(n-1)/2}$ | - Scale independent | - Inability to detect systematic error<br>- Limited to two time points<br>- Not a measure agreement (accuracy) |
| Standard error of measurement (SEM) | Absolute | $Standard\ deviation * \sqrt{(1-ICC)}$ | - Contextualizes each measurement | - Scale dependent |
| Bland-Altman Limits of Agreement (LoA) | Absolute | $\pm t_{0.975,df} * SEM * \sqrt{2}$ ,<br>$Mean\ difference \pm 2 * SD\ of\ differences$ | - Appropriate for methods comparison<br>- Contextualizes each measurement | - Limited to two time points<br>- Biased depending on degrees of freedom (sample size)<br>- Stringency in detecting meaningful change<br>- Scale dependent |
| Coefficient of variation | Absolute | $\frac{Standard\ deviation}{Mean} * 100$ | - Contextualizes each measurement | - Scale dependent |
| T-test, ANOVA | | | - Captures systematic change | - Comparison of means only (not individual differences) |

**Table S2**. First pass quality checks detailed in quality_check(data) in Self_Regulation_Ontology/selfregulation/utils/data_preparation_utils.py. A subject is considered 'failed' if they fail >3 criteria.

| Type | Criteria | Exceptions |
|---|---|---|
| Response time threshold | > 200 ms/trial | Angling risk task: 0 ms<br>Simple reaction time: 150 ms |
| Accuracy threshold | > 0.6 | Digit span: 0<br>Hierarchical rule: 0<br>Information sampling: 0<br>Probabilistic selection: 0<br>Ravens: 0<br>Shift task: 0<br>Spatial span: 0<br>Tower of London: 0 |
| Missed trials threshold | < 0.25 | Information sampling task: 1<br>Go no go: 1<br>Tower of london: 2 |
| Response threshold | > 0.95 | Angling risk task: NA<br>Columbia card task cold: NA<br>Discount titrate: NA<br>Digit span: NA<br>Go no go: .98<br>Kirby: NA<br>Simple reaction time: NA<br>Spatial span: NA |
| Information sampling task | >2 clicks per trial | |
| Psychological refractory period | Both response times >200 ms/trial<br>Both choice accuracies >0.6 | |
| Tower of London | Not making >2 moves in a problem | |
| Two stage | Both response times >200 ms/trial<br>Both responses >0.95 | |
| Writing task | >100 total words | |

**Table S3.** Demographic items

| | |
|---|---|
| Alcohol | How many drinks containing alcohol do you have on a typical day when you are drinking? |
| | How often do you have a drink containing alcohol? |
| | How often do you have six or more drinks on one occasion? |
| | How often during the last year have you found that you were not able to stop drinking once you had started? |
| | How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session? |
| | How often during the last year have you failed to do what was normally expected from you because of drinking? |
| | How often during the last year have you had a feeling of guilt or remorse after drinking? |
| | How often during the last year have you been unable to remember what happened the night before because you had been drinking? |
| | Have you or someone else been injured as a result of your drinking? |
| | Has a relative or friend or a doctor or another health worker been concerned about your drinking or suggested you cut down? |
| Caffeine | On average, how many cans of caffeinated soda do you have each day? |
| | What is your daily caffeine intake from other sources each day (in mg)? |
| | On average, how many cups of coffee do you have each day? |
| | On average, how many cups of tea do you have each day? |
| Drugs | Do you abuse more than one drug at a time? |
| | Are you always able to stop using drugs when you want to? |
| | Have you had "blackouts" or "flashbacks" as a result of drug use? |
| | Have you ever thought about cutting down, or stopping, your use of cannabis? |
| | How many hours were you ‘stoned’ on a typical day when you had been using Cannabis? |
| | How often do you use cannabis? |
| | Have you used any cannabis over the past 6 months? |
| | Have you engaged in illegal activities in order to obtain drugs? |

| | |
|---|---|
| | Do you ever feel bad or guilty about your drug use? |
| | How often during the past six months did you find that you were not able to stop using cannabis once you had started? |
| | How often in the past six months have you devoted a great deal of your time to getting, using, or recovering from cannabis? |
| | How often during the past six months did you fail to do what was normally expected from you because of using cannabis? |
| | How often do you use cannabis in situations that could be physically hazardous, such as driving, operating machinery, or caring for children? |
| | How often in the past six months have you had a problem with your memory or concentration after using cannabis? |
| | Have you had medical problems as a result of your drug use? |
| | Have you neglected your family because of your use of drugs? |
| | Have you used drugs other than those required for medical reasons? |
| | Does your spouse (or parents) ever complain about your involvement with drugs? |
| | Have you ever experienced withdrawal symptoms (felt sick) when you stopped taking drugs? |
| Financial | How much credit card debt do you have? |
| | How much car-related debt do you have? |
| | How much education debt do you have? |
| | What is your household's annual income (in dollars)? |
| | How much mortgage debt do you have? |
| | If you listed any other sources of debt, how much debt do you have? |
| | What is your race? |
| | Do you rent or own your home/apartment? |
| | Do you have a retirement account? |
| | If you do have a retirement account what percent is in stocks? |
| Mental health | During the past 30 days, about how often did you feel … nervous? … hopeless? … restless or fidgety? … so depressed that nothing could cheer you up? … so depressed that nothing could cheer you up? |

| | |
|---|---|
| | The last six questions asked about feelings that might have occurred during the past 30 days. Taking them altogether, did these feelings occur more often in the past 30 days than is usual for you, about the same as usual, or less often than usual? |
| | During the past 30 days, how many days out of 30 were you totally unable to work or carry out your normal activities because of these feelings? |
| | Not counting the days you reported in response to Q3, how many days in the past 30 were you able to do only half or less of what you would normally have been able to do, because of these feelings? |
| | During the past 30 days, how many times did you see a doctor or other health professional about these feelings? |
| | During the past 30 days, how often have physical health problems been the main cause of these feelings? |
| | Do you have or have you ever been diagnosed with any of the following medical conditions (check all that apply)? ADHD, Alcohol Dependency, Anorexia Nervosa, Anxiety Disorder, Autism Spectrum Disorder, Borderline Personality Disorder, Bulimia, Drug Dependency, Depression, Manic-Depressive (Bipolar) illness, Obsessive Compulsive Disorder, Schizophrenia, Other |
| Physical | How old are you (in years)? |
| | How tall are you (in inches: one foot = 12 inches)? |
| | Are you of Hispanic, Latino or Spanish origin? |
| | What is your sex? |
| | How much do you weigh (in pounds)? |
| | Have you been diagnosed with any neurological disorder (e.g. Alzheimer's, Parkinson's)? |
| | Do you have or have you ever been diagnosed with any of the following psychological disorders ? Type II diabetes, Metabolic Syndrome, High Blood Pressure, Heart Disease, Stroke, Cancer, Sleep Apnea, Other |
| Risk taking | How many times in your life have you been arrested and/or charged with illegal activities? |
| | Do you feel you have a problem with gambling? |
| | How many traffic accidents have you been in over your life? |
| | How many traffic tickets have you gotten in the last year? |
| Smoking | On average, how many cigarettes do you now smoke a day (1 pack = 20 cigarettes)? |
| | How long have you smoked (cumulatively)? |
| | How soon after you wake up do you smoke your first cigarette? |
| | Altogether, have you smoked at least 100 or more cigarettes in your entire lifetime? |

| | |
|---|---|
| | In the past 30 days, what tobacco products OTHER THAN cigarettes have you used? |
| | Do you now smoke cigarettes every day, some days or not at all? |
| Social | How many children do you have? |
| | How many times have you been divorced? |
| | What is the highest level of education you have completed? |
| | How long was/is your longest romantic relationship? |
| | What are you motivations for participating in this experiment? |
| | How many romantic relationships have you had? |
| | What is your relationship status? |

**References**


1. Weir JP (2005) Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 19(1):231.

2. Hopkins WG (2000) Measures of reliability in sports medicine and science. *Sports Med* 30(1):1–15.

3. Bruton A, Conway JH, Holgate ST (2000) Reliability: what is it, and how is it measured? *Physiotherapy* 86(2):94–99.

4. Jaeggi SM, et al. (2010) The relationship between n-back performance and matrix reasoning—implications for training and transfer. *Intelligence* 38(6):625–635.

5. Au J, et al. (2015) Improving fluid intelligence with training on working memory: a meta-analysis. *Psychon Bull Rev* 22(2):366–377.

6. Hockey A, Geffen G (2004) The concurrent validity and test--retest reliability of a visuospatial working memory task. *Intelligence* 32(6):591–605.

7. van Leeuwen M, van den Berg SM, Hoekstra RA, Boomsma DI (2007) Endophenotypes for intelligence in children and adolescents. *Intelligence* 35(4):369–380.

8. Studer-Luethi B, Jaeggi SM, Buschkuehl M, Perrig WJ (2012) Influence of neuroticism and conscientiousness on working memory training outcome. *Pers Individ Dif* 53(1):44–49.

9. Pleskac TJ (2008) Decision making and learning while taking sequential risks. *J Exp Psychol Learn Mem Cogn* 34(1):167–185.

10. Lejuez CW, et al. (2002) Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *J Exp Psychol Appl* 8(2):75.

11. Lejuez CW, et al. (2003) The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Exp Clin Psychopharmacol* 11(1):26.

12. White TL, Lejuez CW, de Wit H (2008) Test-retest characteristics of the Balloon Analogue Risk Task (BART). *Exp Clin Psychopharmacol* 16(6):565.

13. Weafer J, Baggott MJ, de Wit H (2013) Test--retest reliability of behavioral measures of impulsive choice, impulsive action, and inattention. *Exp Clin Psychopharmacol* 21(6):475.

14. MacPherson L, Magidson JF, Reynolds EK, Kahler CW, Lejuez CW (2010) Changes in sensation seeking and risk-taking propensity predict increases in alcohol use among early adolescents. *Alcohol Clin Exp Res* 34(8):1400–1408.

15. Fan J, McCandliss BD, Sommer T, Raz A, Posner MI (2002) Testing the efficiency and independence of attentional networks. *J Cogn Neurosci* 14(3):340–347.

16. Fan J, Wu Y, Fossella JA, Posner MI (2001) Assessing the heritability of attentional networks. *BMC Neurosci* 2(1):14.

17. Ishigami Y, Klein RM (2010) Repeated measurement of the components of attention using two versions of the Attention Network Test (ANT): Stability, isolability, robustness, and reliability. *J Neurosci Methods* 190(1):117–128.

18. Habekost T, Petersen A, Vangkilde S (2014) Testing attention: comparing the ANT with TVA-based assessment. *Behav Res Methods* 46(1):81–94.

19. Paap KR, Sawi O (2016) The role of test-retest reliability in measuring individual and group differences in executive functioning. *J Neurosci Methods* 274:81–93.

20. Donders FC (1969) On the speed of mental processes. *Acta Psychol* 30:412–431.

21. Barrett GV, Alexander RA, Doverspike D, Cellar D, Thomas JC (1982) The development and application of a computerized information-processing test battery. *Appl Psychol Meas* 6(1):13–29.

22. Kennedy RS, Baltzley DR, Wilkes RL, Kuntz LA (1989) Psychology of computer use: IX. A menu of self-administered microcomputer-based neurotoxicology tests. *Percept Mot Skills* 68(3_suppl):1255–1272.

23. Williams LM, et al. (2005) The test-retest reliability of a standardized neurocognitive and neurophysiological test battery:``neuromarker''. *Int J Neurosci* 115(12):1605–1630.

24. Deary IJ, Liewald D, Nissan J (2011) A free, easy-to-use, computer-based simple and four-choice reaction time programme: the Deary-Liewald reaction time task. *Behav Res Methods* 43(1):258–268.

25. Jones SAH, et al. (2016) Measuring the performance of attention networks with the Dalhousie computerized attention battery (DalCAB): Methodology and reliability in healthy adults. *Front Psychol* 7.

26. Frederick S (2005) Cognitive Reflection and Decision Making. *J Econ Perspect* 19(4):25–42.

27. Toplak ME, West RF, Stanovich KE (2014) Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Think Reason* 20(2):147–168.

28. Primi C, Morsanyi K, Chiesi F, Donati MA, Hamilton J (2016) The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *J Behav Decis Mak* 29(5):453–469.

29. Haigh M (2016) Has the standard cognitive reflection test become a victim of its own success? *Adv Cogn Psychol* 12(3):145.

30. Stieger S, Reips U-D (2016) A limitation of the Cognitive Reflection Test: familiarity.

*PeerJ* 4:e2395.

31. Figner B, Mackinlay RJ, Wilkening F, Weber EU (2009) Affective and deliberative processes in risky choice: age differences in risk taking in the Columbia Card Task. *J Exp Psychol Learn Mem Cogn* 35(3):709–730.

32. Koffarnus MN, Bickel WK (2014) A 5-trial adjusting delay discounting task: accurate discount rates in less than one minute. *Exp Clin Psychopharmacol* 22(3):222.

33. Richards JB, Mitchell SH, Wit H, Seiden LS (1997) Determination of discount functions in rats with an adjusting-amount procedure. *J Exp Anal Behav* 67(3):353–366.

34. Richards JB, Zhang L, Mitchell SH, Wit H (1999) Delay or probability discounting in a model of impulsive behavior: effect of alcohol. *J Exp Anal Behav* 71(2):121–143.

35. Baker F, Johnson MW, Bickel WK (2003) Delay discounting in current and never-before cigarette smokers: similarities and differences across commodity, sign, and magnitude. *J Abnorm Psychol* 112(3):382.

36. Johnson MW, Bickel WK, Baker F (2007) Moderate drug use and delay discounting: a comparison of heavy, light, and never smokers. *Exp Clin Psychopharmacol* 15(2):187–194.

37. Reed DD, Martens BK (2011) Temporal discounting predicts student responsiveness to exchange delays in a classroom token system. *J Appl Behav Anal* 44(1):1–18.

38. Smits RR, Stein JS, Johnson PS, Odum AL, Madden GJ (2013) Test--retest reliability and construct validity of the Experiential Discounting Task. *Exp Clin Psychopharmacol* 21(2):155.

39. De Wilde B, Bechara A, Sabbe B, Hulstijn W, Dom G (2013) Risky decision-making but not delay discounting improves during inpatient treatment of polysubstance dependent alcoholics. *Front Psychiatry* 4.

40. Hendrickson KL, Rasmussen EB (2013) Effects of mindful eating training on delay and probability discounting for food and money in obese and healthy-weight individuals. *Behav Res Ther* 51(7):399–409.

41. Yoon H, Chapman GB (2016) A closer look at the yardstick: A new discount rate measure with precision and range. *J Behav Decis Mak* 29(5):470–480.

42. Horan WP, Johnson MW, Green MF (2017) Altered experiential, but not hypothetical, delay discounting in schizophrenia. *J Abnorm Psychol* 126(3):301.

43. McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. *Science* 306(5695):503–507.

44. Kirby KN, Maraković NN (1996) Delay-discounting probabilistic rewards: Rates decrease as amounts increase. *Psychon Bull Rev* 3(1):100–104.

45. Kirby KN (2009) One-year temporal stability of delay-discount rates. *Psychon Bull Rev* 16(3):457–462.

46. Black AC, Rosen MI (2011) A money management-based substance use treatment increases valuation of future rewards. *Addict Behav* 36(1):125–128.

47. Wölbert E, Riedl A (2013) Measuring time and risk preferences: Reliability, stability, domain specificity.

48. Arfer KB, Luhmann CC (2017) Time-preference tests fail to predict behavior related to self-control. *Front Psychol* 8.

49. Hare TA, Camerer CF, Rangel A (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324(5927):646–648.

50. Vlaev I, Crockett MJ, Clark L, Müller U, Robbins TW (2017) Serotonin enhances the impact of health information on food choice. *Cogn Affect Behav Neurosci* 17(3):542–553.

51. Wechsler D (1949) Wechsler Intelligence Scale for Children. Available at: http://psycnet.apa.org/record/1950-02930-000 [Accessed April 18, 2018].

52. Baddeley A, Gardner JM, Grantham-McGregor S (1995) Cross-cultural cognition: Developing tests for developing countries. *Appl Cogn Psychol* 9(7).

53. Karpicke J, Pisoni DB (2000) Memory span and sequence learning using multimodal stimulus patterns: Preliminary findings in normal-hearing adults. *Research on Spoken Language Processing*.

54. Williams RA, et al. (2000) Changes in directed attention and short-term memory in depression. *J Psychiatr Res* 34(3):227–238.

55. Sternberg RJ, et al. (2002) Assessing intellectual potential in rural Tanzanian school children. *Intelligence* 30(2):141–162.

56. Gray S (2003) Diagnostic accuracy and test--retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *J Commun Disord* 36(2):129–151.

57. Alloway TP, Gathercole SE, Willis C, Adams A-M (2004) A structural analysis of working memory and related cognitive skills in young children. *J Exp Child Psychol* 87(2):85–106.

58. Tröster AI, Woods SP, Morgan EE (2007) Assessing cognitive change in Parkinson's disease: development of practice effect-corrected reliable change indices. *Arch Clin Neuropsychol* 22(6):711–718.

59. Alloway TP, Gathercole SE, Pickering SJ (2006) Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Dev* 77(6):1698–1716.

60. Slade PD, et al. (2008) The serial use of child neurocognitive tests: development versus practice effects. *Psychol Assess* 20(4):361.

61. Ven SHG, Kroesbergen EH, Boom J, Leseman PPM (2013) The structure of executive functions in children: A closer examination of inhibition, shifting, and updating. *Br J Dev Psychol* 31(1):70–87.

62. Bjork RA (1970) Positive forgetting: The noninterference of Items intentionally forgotten. *Journal of Verbal Learning and Verbal Behavior* 9(3):255–268.

63. Nee DE, Jonides J, Berman MG (2007) Neural mechanisms of proactive interference-resolution. *Neuroimage* 38(4):740–751.

64. MacDonald AW 3rd, et al. (2005) A convergent-divergent approach to context processing, general intellectual functioning, and the genetic liability to schizophrenia. *Neuropsychology* 19(6):814–821.

65. Rosvold HE, Mirsky AF, Sarason I, Bransome ED Jr, Beck LH (1956) A continuous performance test of brain damage. *J Consult Psychol* 20(5):343.

66. Servan-Schreiber D, Cohen JD, Steingard S (1996) Schizophrenic deficits in the processing of context. A test of a theoretical model. *Arch Gen Psychiatry* 53(12):1105–1112.

67. Jones JAH, Sponheim SR, Waw MA (2010) The dot pattern expectancy task: reliability and replication of deficits in schizophrenia.

68. Strauss ME, et al. (2013) Temporal stability and moderating effects of age and sex on CNTRaCS task performance. *Schizophr Bull* 40(4):835–844.

69. Kertzman S, et al. (2008) Go--no-go performance in pathological gamblers. *Psychiatry Res* 161(1):1–10.

70. Bender AD, Filmer HL, Garner KG, Naughtin CK, Dux PE (2016) On the relationship between response selection and response inhibition: An individual differences approach. *Atten Percept Psychophys* 78(8):2420–2432.

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

72. Badre D, Kayser AS, D'Esposito M (2010) Frontal cortex and the discovery of abstract action rules. *Neuron* 66(2):315–326.

73. Holt CA, Laury SK (2005) Risk aversion and incentive effects: New data without order effects. *Am Econ Rev* 95(3):902–912.

74. Andersen S, Harrison GW, Lau MI, Elisabet Rutström E (2008) Lost in state space: are preferences stable? *Int Econ Rev* 49(3):1091–1112.

75. Lönnqvist J-E, Verkasalo M, Walkowitz G, Wichardt PC (2015) Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *J Econ Behav Organ* 119:254–266.

76. Chung D, et al. (2017) Valuation in major depression is intact and stable in a non-learning environment. *Sci Rep* 7.

77. Clark L, Robbins TW, Ersche KD, Sahakian BJ (2006) Reflection impulsivity in current and former substance users. *Biol Psychiatry* 60(5):515–522.

78. Grummitt J, Gaudreau H, Steiner M, Meaney M, Levitan R Cognitive Development in 4-to 6-Year Old Children: Does the Cambridge Neuropsychological Test Automated Battery (CANTAB) Provide a Reliable Assessment?

79. Yntema DB (1963) Keeping track of several things at once. *Hum Factors* 5:7–17.

80. Navon D (1977) Forest before trees: The precedence of global features in visual perception. *Cogn Psychol* 9(3):353–383.

81. Dale G, Arnell KM (2013) Investigating the stability of and relationships among global/local processing measures. *Atten Percept Psychophys* 75(3):394–406.

82. Frank MJ, Seeberger LC, O'reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306(5703):1940–1943.

83. Baker TE, Stockwell T, Holroyd CB (2013) Constraints on decision making: Implications from genetics, personality, and addiction. *Cogn Affect Behav Neurosci* 13(3):417–436.

84. Pashler H (1994) Dual-task interference in simple tasks: data and theory. *Psychol Bull* 116(2):220–244.

85. Welford AT (1952) The "psychological refractory period"and the timing of high-speed performance—a review and a theory. *Br J Psychol* 43(1):2–19.

86. Telford CW (1931) The refractory phase of voluntary and associative responses. *J Exp Psychol* 14(1):1.

87. Raven JC (1948) The comparative assessment of intellectual ability. *Br J Psychol* 39(1):12–19.

88. Watts K, Baddeley A, Williams M (1982) Automated tailored testing using Raven's Matrices and the Mill Hill Vocabulary tests: a comparison with manual administration. *Int J Man Mach Stud* 17(3):331–344.

89. Calvert EJ, Waterfall RC (1982) A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. *Int J Man Mach Stud* 17(3):305–310.

90. Bors DA, Stokes TL (1998) Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educ Psychol Meas* 58(3):382–398.

91. Choudhury N, Gorman KS (1999) The relationship between reaction time and psychometric intelligence in a rural Guatemalan adolescent population. *Int J Psychol* 34(4):209–217.

92. Arthur W Jr, Tubre TC, Paul DS, Sanchez-Ku ML (1999) College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *J Psychoeduc Assess* 17(4):354–361.

93. Costenbader V, Ngari SM (2001) A Kenya standardization of the Raven's coloured progressive matrices. *Sch Psychol Int* 22(3):258–268.

94. Bors DA, Vigneau F (2003) The effect of practice on Raven's Advanced Progressive Matrices. *Learn Individ Differ* 13(4):291–312.

95. Williams JE, McCord DM (2006) Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Comput Human Behav* 22(5):791–800.

96. Sternberg S (1966) High-speed scanning in human memory. *Science* 153(3736):652–654.

97. Monsell S (1978) Recency, immediate recognition memory, and reaction time. *Cogn Psychol* 10(4):465–501.

98. Barch DM, et al. (2008) CNTRICS final task selection: working memory. *Schizophr Bull* 35(1):136–152.

99. DeSchepper B, Treisman A (1996) Visual memory for novel shapes: implicit coding without attention. *J Exp Psychol Learn Mem Cogn* 22(1):27.

100. Wilson RC, Niv Y (2011) Inferring relevance in a changing world. *Front Hum Neurosci* 5:189.

101. Paolo AM, Axelrod BN, Tröster AI (1996) Test-retest stability of the Wisconsin Card Sorting Test. *Assessment* 3(2):137–143.

102. Ingram F, Greve KW, Ingram PTF, Soukup VM (1999) Temporal stability of the Wisconsin Card Sorting Test in an untreated patient sample. *British Journal of Clinical Psychology* 38(2):209–211.

103. Greve KW, et al. (2002) Temporal stability of the Wisconsin Card Sorting Test in a chronic traumatic brain injury sample. *Assessment* 9(3):271–277.

104. Bird CM, Papadopoulou K, Ricciardelli P, Rossor MN, Cipolotti L (2004) Monitoring cognitive changes: Psychometric properties of six cognitive tests. *British Journal of Clinical Psychology* 43(2):197–210.

105. Simon JR, Rudell AP (1967) Auditory SR compatibility: the effect of an irrelevant cue on information processing. *J Appl Psychol* 51(3):300.

106. de Jong P, van den Hout M, Rietbroek H, Huijding J (2003) Dissociations between implicit and explicit attitudes toward phobic stimuli. *Cogn Emot* 17(4):521–545.

107. Wöstmann NM, et al. (2013) Reliability and plasticity of response inhibition and interference control. *Brain Cogn* 81(1):82–94.

108. Linck JA, Weiss DJ (2015) Can working memory and inhibitory control predict second language learning in the classroom? *SAGE Open* 5(4):2158244015607352.

109. Collie A, Maruff P, Darby DG, McSTEPHEN M (2003) The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test--retest intervals. *J Int Neuropsychol Soc* 9(3):419–428.

110. Erlanger D, et al. (2003) Development and validation of a web-based neuropsychological test protocol for sports-related return-to-play decision-making. *Arch Clin Neuropsychol* 18(3):293–316.

111. Lemay S, Bédard M-A, Rouleau I, Tremblay P-L (2004) Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *Clin Neuropsychol* 18(2):284–302.

112. Falleti MG, Maruff P, Collie A, Darby DG (2006) Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *J Clin Exp Neuropsychol* 28(7):1095–1112.

113. Cole WR, et al. (2013) Test--retest reliability of four computerized neurocognitive assessment tools in an active duty military population. *Arch Clin Neuropsychol* 28(7):732–742.

114. Corsi P (1972) Memory and the medial temporal region of the brain. *Unpublished doctoral dissertation), McGill University, Montreal, QB*. Available at: http://digitool.library.mcgill.ca/webclient/DeliveryManager?pid=70754&custom_att_2=direct.

115. Orsini A (1994) Corsi's Block-Tapping Test: Standardization and Concurrent Validity with WISC---R for Children Aged 11 to 16. *Percept Mot Skills* 79(3_suppl):1547–1554.

116. Lowe C, Rabbitt P (1998) Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. *Neuropsychologia* 36(9):915–923.

117. Cho B, et al. (2002) The validity and reliability of a Computerized Dementia Screening Test developed in Korea. *J Neurol Sci* 203:109–114.

118. Saggino A, Balsamo M, Grieco A, Cerbone MR, Raviele NN (2004) Corsi's Block-

Tapping Task: Standardization and Location in Factor Space with the Wais--R for Two Normal Samples of Older Adults. *Percept Mot Skills* 98(3):840–848.

119. Fisher A, Boyle J, Paton JY (2011) Effects of a physical education intervention on cognitive function in young children: randomized controlled pilot study. *Biomed Chromatogr*. Available at: https://bmcpediatr.biomedcentral.com/articles/10.1186/1471-2431-11-97.

120. Lo AHY, Humphreys M, Byrne GJ, Pachana NA (2012) Test--retest reliability and practice effects of the Wechsler Memory Scale-III. *J Neuropsychol* 6(2):212–231.

121. Wechsler D (1997) *Wechsler Adult Intelligence Scale--Third Edition (WAIS--III)* (San Antonio, TX: The Psychological Corporation).

122. Lappin JS, Eriksen CW (1966) Use of a delayed signal to stop a visual reaction-time response. *J Exp Psychol* 72(6):805.

123. Logan GD, Cowan WB (1984) On the ability to inhibit thought and action: A theory of an act of control. *Psychol Rev* 91(3):295.

124. Vince MA (1948) The intermittency of control movements and the psychological refractory period. *Br J Psychol Gen Sect* 38(Pt 3):149–157.

125. Kindlon D, Mezzacappa E, Earls F (1995) Psychometric properties of impulsivity measures: Temporal stability, validity and factor structure. *J Child Psychol Psychiatry* 36(4):645–661.

126. Kuntsi J, Stevenson J, Oosterlaan J, Sonuga-Barke EJS (2001) Test-retest reliability of a new delay aversion task and executive function measures. *Br J Dev Psychol* 19(3):339–348.

127. Soreni N, Crosbie J, Ickowicz A, Schachar R (2009) Stop signal and conners' continuous performance tasks: Test---retest reliability of two inhibition measures in adhd children. *J Atten Disord* 13(2):137–143.

128. De Jong R, Coles MG, Logan GD (1995) Strategies and mechanisms in nonselective and selective inhibitory motor control. *J Exp Psychol Hum Percept Perform* 21(3):498–511.

129. Bissett PG, Logan GD (2014) Selective stopping? Maybe not. *J Exp Psychol Gen* 143(1):455–472.

130. Bedard A-C, et al. (2002) The development of selective inhibitory control across the life span. *Dev Neuropsychol* 21(1):93–111.

131. Stroop JR (1935) Studies of interference in serial verbal reactions. *J Exp Psychol* 18(6):643.

132. Franzen MD, Tishelman AC, Sharp BH, Friedman AG (1987) An investigation of the test-retest reliability of the stroop colorword test across two intervals. *Arch Clin*

*Neuropsychol* 2(3):265–272.

133. Harbeson MM, Krause M, Kennedy RS, Bittner AC Jr (1982) The Stroop as a performance evaluation test for environmental research. *J Psychol* 111(2):223–233.

134. Siegrist M (1997) Test-retest reliability of different versions of the Stroop test. *J Psychol* 131(3):299–306.

135. Strauss GP, Allen DN, Jorgensen ML, Cramer SL (2005) Test-retest reliability of standard and emotional stroop tasks: an investigation of color-word and picture-word versions. *Assessment* 12(3):330–337.

136. Logan GD, Bundesen C (2003) Clever homunculus: is there an endogenous act of control in the explicit task-cuing procedure? *J Exp Psychol Hum Percept Perform* 29(3):575–599.

137. Mayr U, Kliegl R (2003) Differential effects of cue changes and task changes on task-set selection costs. *J Exp Psychol Learn Mem Cogn* 29(3):362.

138. Shallice T (1982) Specific impairments of planning. *Philos Trans R Soc Lond B Biol Sci* 298(1089):199–209.

139. Schnirman GM, Welsh MC, Retzlaff PD (1998) Development of the Tower of London-revised. *Assessment* 5(4):355–360.

140. Keefe RSE, et al. (2004) The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophr Res* 68(2):283–297.

141. Kaneda Y, et al. (2007) Brief assessment of cognition in schizophrenia: validation of the Japanese version. *Psychiatry Clin Neurosci* 61(6):602–609.

142. Dockery CA, Hueckel-Weng R, Birbaumer N, Plewnia C (2009) Enhancement of planning ability by transcranial direct current stimulation. *Journal of Neuroscience* 29(22):7271–7277.

143. Bouso JC, Fábregas JM, Antonijoan RM, Rodríguez-Fornells A, Riba J (2013) Acute effects of ayahuasca on neuropsychological performance: differences in executive function between experienced and occasional users. *Psychopharmacology* 230(3):415–424.

144. Köstering L, Nitschke K, Schumacher FK, Weiller C, Kaller CP (2015) Test--retest reliability of the Tower of London Planning Task (TOL-F). *Psychol Assess* 27(3):925.

145. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69(6):1204–1215.

146. Patton JH, Stanford MS, Barratt ES (1995) Factor structure of the Barratt impulsiveness scale. *J Clin Psychol* 51(6):768–774.

147. Stanford MS, et al. (2009) Fifty years of the Barratt Impulsiveness Scale: An update and review. *Pers Individ Dif* 47(5):385–395.

148. Fossati A, Di Ceglie A, Acquarini E, Barratt ES (2001) Psychometric properties of an Italian version of the Barratt Impulsiveness Scale-11 (BIS-11) in nonclinical subjects. *J Clin Psychol* 57(6):815–828.

149. Surís A, Borman PD, Lind L, Kashner TM (2007) Aggression, impulsivity, and health functioning in a veteran population: equivalency and test--retest reliability of computerized and paper-and-pencil administrations. *Comput Human Behav* 23(1):97–110.

150. Güleç H, et al. (2008) Psychometric Properties of the Turkish Version of the Barratt Impulsiveness Scale-11. *Klinik Psikofarmakoloji Bulteni* 18(4).

151. Tangney JP, Baumeister RF, Boone AL (2004) High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *J Pers* 72(2):271–324.

152. Maloney PW, Grawitch MJ, Barber LK (2012) The multi-factor structure of the Brief Self-Control Scale: Discriminant validity of restraint and impulsivity. *J Res Pers* 46(1):111–115.

153. Duckworth AL, Seligman MEP (2005) Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychol Sci* 16(12):939–944.

154. Dickman SJ (1990) Functional and dysfunctional impulsivity: personality and cognitive correlates. *J Pers Soc Psychol* 58(1):95–102.

155. Brunas-Wagstaff J, Tilley A, Verity M, Ford S, Thompson D (1997) Functional and dysfunctional impulsivity in children and their relationship to Eysenck's impulsiveness and venturesomeness dimensions. *Pers Individ Dif* 22(1):19–25.

156. Caci H, Nadalet L, Baylé FJ, Robert P, Boyer P (2003) Functional and dysfunctional impulsivity: contribution to the construct validity. *Acta Psychiatr Scand* 107(1):34–40.

157. Chico E, Tous JM, Lorenzo-Seva U, Vigil-Colet A (2003) Spanish adaptation of Dickman's impulsivity inventory: its relationship to Eysenck's personality questionnaire. *Pers Individ Dif* 35(8):1883–1892.

158. Weber EU, Blais A-R, Betz NE (2002) A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *J Behav Decis Mak* 15(4):263–290.

159. Blais A-R, Weber EU (2006) A domain-specific risk-taking (DOSPERT) scale for adult populations.

160. Du X, Li J, Du X (2014) Testing risk-taking behavior in Chinese undergraduate students. *PLoS One* 9(5):e97989.

161. Gross JJ, John OP (2003) Individual differences in two emotion regulation processes:

implications for affect, relationships, and well-being. *J Pers Soc Psychol* 85(2):348.

162. Balzarotti S, John OP, Gross JJ (2010) An Italian adaptation of the emotion regulation questionnaire. *Eur J Psychol Assess*.

163. Gullone E, Taffe J (2012) The Emotion Regulation Questionnaire for Children and Adolescents (ERQ--CA): A psychometric evaluation. *Psychol Assess* 24(2):409.

164. Cabello R, Salguero JM, Fernández-Berrocal P, Gross JJ (2013) A Spanish adaptation of the emotion regulation questionnaire. *Eur J Psychol Assess*.

165. Eldeleklioğlu J, Eroğlu Y (2015) A Turkish adaptation of the Emotion Regulation Questionnaire. *Journal of Human Sciences* 12(1):1157–1168.

166. Hasani J (2017) Persian Version of the Emotion Regulation Questionnaire: Factor Structure, Reliability and Validity. *International Journal of Behavioral Sciences* 10(4):156–161.

167. Baer RA, Smith GT, Hopkins J, Krietemeyer J, Toney L (2006) Using self-report assessment methods to explore facets of mindfulness. *Assessment* 13(1):27–45.

168. Isenberg L (2009) Mindfulness--Life with attention and awarenes: Test-retest reliability of the FFMQ for Dutch fibromyalgia patients. Dissertation (University of Twente).

169. Deng Y-Q, Liu X-H, Rodriguez MA, Xia C-Y (2011) The five facet mindfulness questionnaire: Psychometric properties of the Chinese version. *Mindfulness* 2(2):123–128.

170. Petrocchi N, Ottaviani C (2016) Mindfulness facets distinctively predict depressive symptoms after two years: The mediating role of rumination. *Pers Individ Dif* 93:92–96.

171. Carstensen LL, Lang FR (1996) Future orientation scale. *Unpublished manuscript, Stanford University*.

172. Zacher H, de Lange AH (2011) Relations between chronic regulatory focus and future time perspective: Results of a cross-lagged structural equation model. *Pers Individ Dif* 50(8):1255–1260.

173. Kooij DT, Bal PM, Kanfer R (2014) Future time perspective and promotion focus as determinants of intraindividual change in work motivation. *Psychol Aging* 29(2):319.

174. Duckworth AL, Quinn PD (2009) Development and validation of the Short Grit Scale (GRIT--S). *J Pers Assess* 91(2):166–174.

175. Hill PL, Burrow AL, Bronk KC (2016) Persevering with positivity and purpose: An examination of purpose commitment and positive affect as predictors of grit. *J Happiness Stud* 17(1):257–269.

176. Li J, et al. (2016) Psychometric assessment of the Short Grit Scale among Chinese

adolescents. *J Psychoeduc Assess*:0734282916674858.

177. Eysenck SBG, Pearson PR, Easting G, Allsopp JF (1985) Age norms for impulsiveness, venturesomeness and empathy in adults. *Pers Individ Dif* 6(5):613–619.

178. Luengo MA, Carrillo-De-La-Pena MT, Otero JM (1991) The components of impulsiveness: A comparison of the I. 7 Impulsiveness Questionnaire and the Barratt Impulsiveness Scale. *Pers Individ Dif* 12(7):657–667.

179. Brown KW, Ryan RM (2003) The benefits of being present: mindfulness and its role in psychological well-being. *J Pers Soc Psychol* 84(4):822.

180. Barnes S, Brown KW, Krusemark E, Campbell WK, Rogge RD (2007) The role of mindfulness in romantic relationship satisfaction and responses to relationship stress. *J Marital Fam Ther* 33(4):482–500.

181. Deng Y-Q, et al. (2012) Psychometric properties of the Chinese translation of the Mindful Attention Awareness Scale (MAAS). *Mindfulness* 3(1):10–14.

182. Murphy MJ, Mermelstein LC, Edwards KM, Gidycz CA (2012) The benefits of dispositional mindfulness in physical health: a longitudinal study of female college students. *J Am Coll Health* 60(5):341–348.

183. Matvienko-Sikar K, Dockray S (2017) Effects of a novel positive psychological intervention on prenatal stress and well-being: A pilot randomised controlled trial. *Women Birth* 30(2):e111–e118.

184. Whiteside SP, Lynam DR (2001) The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Pers Individ Dif* 30(4):669–689.

185. Baltes PB, Staudinger UM, Lindenberger U (1999) Lifespan psychology: Theory and application to intellectual functioning. *Annu Rev Psychol* 50(1):471–507.

186. Freund AM, Baltes PB (2002) Life-management strategies of selection, optimization and compensation: Measurement by self-report and construct validity. *J Pers Soc Psychol* 82(4):642.

187. Zuckerman M, Kolin EA, Price L, Zoob I (1964) Development of a sensation-seeking scale. *J Consult Psychol* 28(6):477.

188. Zuckerman M (1971) Dimensions of sensation seeking. *J Consult Clin Psychol* 36(1):45.

189. Zuckerman M, Bone RN, Neary R, Mangelsdorff D, Brustman B (1972) What is the sensation seeker? Personality trait and experience correlates of the Sensation-Seeking Scales. *J Consult Clin Psychol* 39(2):308.

190. Zaleski Z (1984) Sensation-seeking and risk-taking behaviour. *Pers Individ Dif* 5(5):607–608.

191. Thombs DL, Beck KH, Mahoney CA, Bromley MD, Bezon KM (1994) Social Context, Sensation Seeking, and Teen-age Alcohol Abuse. *J Sch Health* 64(2):73–79.

192. Parmak M, Mylle JJC, Euwema MC (2013) Personality and the perception of situation structure in a military environment: seeking sensation versus structure as a soldier. *J Appl Soc Psychol* 43(5):1040–1049.

193. Brown JM, Miller WR, Lawendowski LA (1999) The self-regulation questionnaire.

194. Carey KB, Neal DJ, Collins SE (2004) A psychometric analysis of the self-regulation questionnaire. *Addict Behav* 29(2):253–260.

195. Neal DJ, Carey KB (2005) A follow-up psychometric analysis of the self-regulation questionnaire. *Psychol Addict Behav* 19(4):414.

196. Kiernan M, et al. (2013) The Stanford Leisure-Time Activity Categorical Item (L-Cat): a single categorical item sensitive to physical activity changes in overweight/obese women. *Int J Obes* 37(12):1597.

197. Riebl SK, et al. (2013) The comparative validity of interactive multimedia questionnaires to paper-administered questionnaires for beverage intake and physical activity: pilot study. *JMIR Res Protoc* 2(2).

198. Gosling SD, Rentfrow PJ, Swann WB (2003) A very brief measure of the Big-Five personality domains. *J Res Pers* 37(6):504–528.

199. Romero E, Villar P, Gómez-Fraguela JA, López-Romero L (2012) Measuring personality traits with ultra-short scales: A study of the Ten Item Personality Inventory (TIPI) in a Spanish sample. *Pers Individ Dif* 53(3):289–293.

200. Renau V, Oberst Ú, Gosling SD, Rusiñol J, Lusar AC (2013) Translation and validation of the ten-item-personality inventory into Spanish and Catalan. *Aloma: revista de psicologia, ciències de l'educació i de l'esport Blanquerna* (31 (2)):85–97.

201. Job V, Dweck CS, Walton GM (2010) Ego depletion---Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychol Sci* 21(11):1686–1693.

202. Karlsson J, Persson L-O, Sjöström L, Sullivan M (2000) Psychometric properties and factor structure of the Three-Factor Eating Questionnaire (TFEQ) in obese men and women. Results from the Swedish Obese Subjects (SOS) study. *Int J Obes* 24(12):1715.

203. Mostafavi S-A, et al. (2017) The Reliability and validity of the Persian Version of Three-Factor Eating Questionnaire-R18 (TFEQ-R18) in Overweight and Obese Females. *Iran J Psychiatry* 12(2):100.

204. Lynam DR, Smith GT, Whiteside SP, Cyders MA (2006) The UPPS-P: Assessing five personality pathways to impulsive behavior. *West Lafayette, IN: Purdue University*.

205. Anestis MD, Selby EA, Joiner TE (2007) The role of urgency in maladaptive behaviors. *Behav Res Ther* 45(12):3018–3029.

206. Kaiser AJ (2015) *Bidirectional Relations of Impulsive Personality and Alcohol Use Over Two Years* (University of Kentucky).

207. Zimbardo PG, Boyd JN (1999) Putting Time in Perspective: A Valid, Reliable Individual-Differences Metric. *J Pers Soc Psychol* 77(6):1271–1288.

208. Kolesovs A (2009) Factorial validity of the Latvian and Russian versions of the Zimbardo Time Perspective Inventory in Latvia. *Baltic Journal of Psychology* 10(1-2):55–64.

209. Carelli MG, Wiberg B, Wiberg M (2011) Development and construct validation of the Swedish Zimbardo time perspective inventory. *Eur J Psychol Assess*.

210. Zhang JW, Howell RT, Bowerman T (2013) Validating a brief measure of the Zimbardo Time Perspective Inventory. *Time & Society* 22(3):391–409.

211. Wang Y, Chen X-J, Cui J-F, Liu L-L (2015) Testing the Zimbardo time perspective inventory in the Chinese context. *PsyCh journal* 4(3):166–175.