Title: Uncovering mental structure through data-driven ontology discovery

One Sentence Summary: A framework for cumulative psychological science is used to distill psychological theory and evaluate predictive ability.

Authors: Ian W Eisenberg[1,*], Patrick G Bissett[1], A Zeynep Enkavi[1], Jamie Li[1], David P. MacKinnon[3], Lisa A. Marsch[2,], Russell A. Poldrack[1]

Affiliations:
[1] Department of Psychology, Stanford University, Stanford, California 94305
[2] Department of Psychiatry, Geisel School of Medicine at Dartmouth, Dartmouth College, Lebanon, NH 03766
[3] Department of Psychology, Arizona State University, Tempe, AZ 85281
*Corresponding Author: ieisenbe@stanford.edu

Abstract: Psychological sciences have identified a wealth of cognitive processes and behavioral phenomena, yet struggle to produce cumulative knowledge. Progress is hamstrung by siloed scientific traditions and a focus on explanation over prediction, two issues we address by examining individual differences across an unprecedented range of behavioral tasks, self-report surveys, and real-world outcomes. We derive a cognitive ontology and evaluate the predictive power of many psychological measurements related to self-regulation. Though both tasks and surveys putatively measure self-regulation, they show little empirical relationship. Within tasks and surveys, however, the ontology reveals opportunities for theoretic synthesis and identifies stable individual traits. Additionally, surveys predict self-reported real-world outcomes while tasks largely do not. We conclude that data-driven ontologies lay the groundwork for a cumulative psychological science.

Main Text:

Science is meant to be cumulative, but both methodological and conceptual problems have impeded cumulative progress in psychological science. While a flurry of recent work has focused on the poor reproducibility of psychological findings (*1*), a more fundamental conceptual challenge arises from the lack of integrative theory development and testing. As pointed out by Newell (*2*) and Meehl (*3*) decades ago, psychological findings are rarely contextualized within the broader literature and the resulting theories are siloed and overspecialized. Thus it seems essential to develop an integrative framework, one that capitalizes on the wealth of psychological phenomena to create a foundation for future inquiry (*4*). We propose that the data-driven development of ontologies - formal descriptions of concepts in a domain and their relationships (*5*)- can serve as such a framework. By specifying latent psychological constructs and their relationship to specific measures, ontologies can serve as a *lingua franca* between literatures, bridge disciplines, identify theoretical gaps, and clarify research programs (*6*). In this paper, we integrate a large array of psychological measures into an ontological framework, via a large-scale study of behavioral individual differences.

Theoretic integration and construct validity (*7*, *8*) should be complemented by ecological validity. The constructs studied by psychologists are hypothesized to serve as building blocks for everyday behavior, and their dysfunction is thought to be central to many disorders of mental health (*9*). However, these constructs are often derived to explain behavior in an ad hoc manner, rather than to generate a priori predictions of real-world behavior, leaving this link largely untested (*10*). Even when associations between psychological constructs and real-world behavior are examined, they rarely are evaluated using modern assessments of predictive accuracy (e.g., (*11*), leading to generally inflated estimates of predictive power (*12*). We evaluate the ability of psychological measurements to predict a range of real-world behaviors, and unpack the predictive success in terms of the ontology. Linking disparate real-world outcomes based on ontological similarity is a critical step towards creating a contextualized, generalizable science of human behavior.

Ontology creation

Cognitive ontologies imply a similarity structure between psychological measurements, thus their structure can be inferred from statistical techniques that capitalize on similarity between variables. This strategy derives from a classic approach in psychology, factor analysis, which has been used to infer the structure of broad constructs such as personality (*13*, *14*) and emotion (*15*), inform measurement design (*16*), and drive hypothesis development (*17*). However, these previous studies have suffered from a limited measurement scope, restricting their integrative capacity. Exceptions that evaluate many measures do exist, but they have tended towards confirmatory analyses, distinguishing between particular conceptualizations of mental processes (*18*), (*19*), and have not explored relationships across measurement categories (e.g. self-report vs. behavioral tasks). To create a holistic cognitive ontology, it is necessary to broaden the scope, both of the behavioral measurements and of the hypothesis space implicit in the analytic strategy.

We applied the data-driven ontology development approach within the psychological domain of self-regulation, which refers to the ability to regulate behavior in service of longer-term goals. This domain is an ideal starting point for ontological revision due to its debated multi-faceted nature (*20*), putative connection to real-world behaviors (*21*), and measurement diversity, involving self-report surveys and behavioral tasks measuring performance (*22*, *23*). We selected a set of 23 self-report surveys (Table S1) and 37 behavioral tasks (Table S2) in order to capture relevant constructs in this domain (e.g., temporal discounting, cognitive control and impulsivity) while also including a broader set of measures to capture a diverse psychological space that extends beyond those normally studied in the context of self-regulation (e.g. information processing, personality). We acquired data for this entire set of measures (~10 hours of data collection) from 522 participants. A subgroup of 150 participants completed a retest on the entire battery (60 - 228 days after initial test), allowing estimation of retest reliability (Figure S1).

Once selected, behavior on each of these 60 measures was decomposed into multiple dependent variables (DVs; N=196; see Table S1-2) which reflect means of specific item sets, comparisons between task conditions, or model parameters thought to capture psychological constructs (Figure 1a,b). For example, the task-switching task is a measure that is decomposed into multiple DVs including task-switch cost and cue-switch cost (*24*, *25*). Where appropriate, reaction time and accuracy data on two-choice tasks were modeled using the drift diffusion model (*26*), and model parameters were used as DVs. The resulting DVs are the fundamental unit of measurement in our analyses.

In addition to supporting the specific goals of this project, the experimental code (*27*), raw data, data cleaning procedures, and analysis code were designed to be disseminated and used openly and are being released alongside this paper. The data acquisition plan was pre-registered on the Open Science Framework (http://goo.gl/3eJuu1); subsequent analyses should be considered exploratory. See the supplement for a full description of data acquisition, quality assurance procedures, replication efforts, and the full list of measures and DVs.

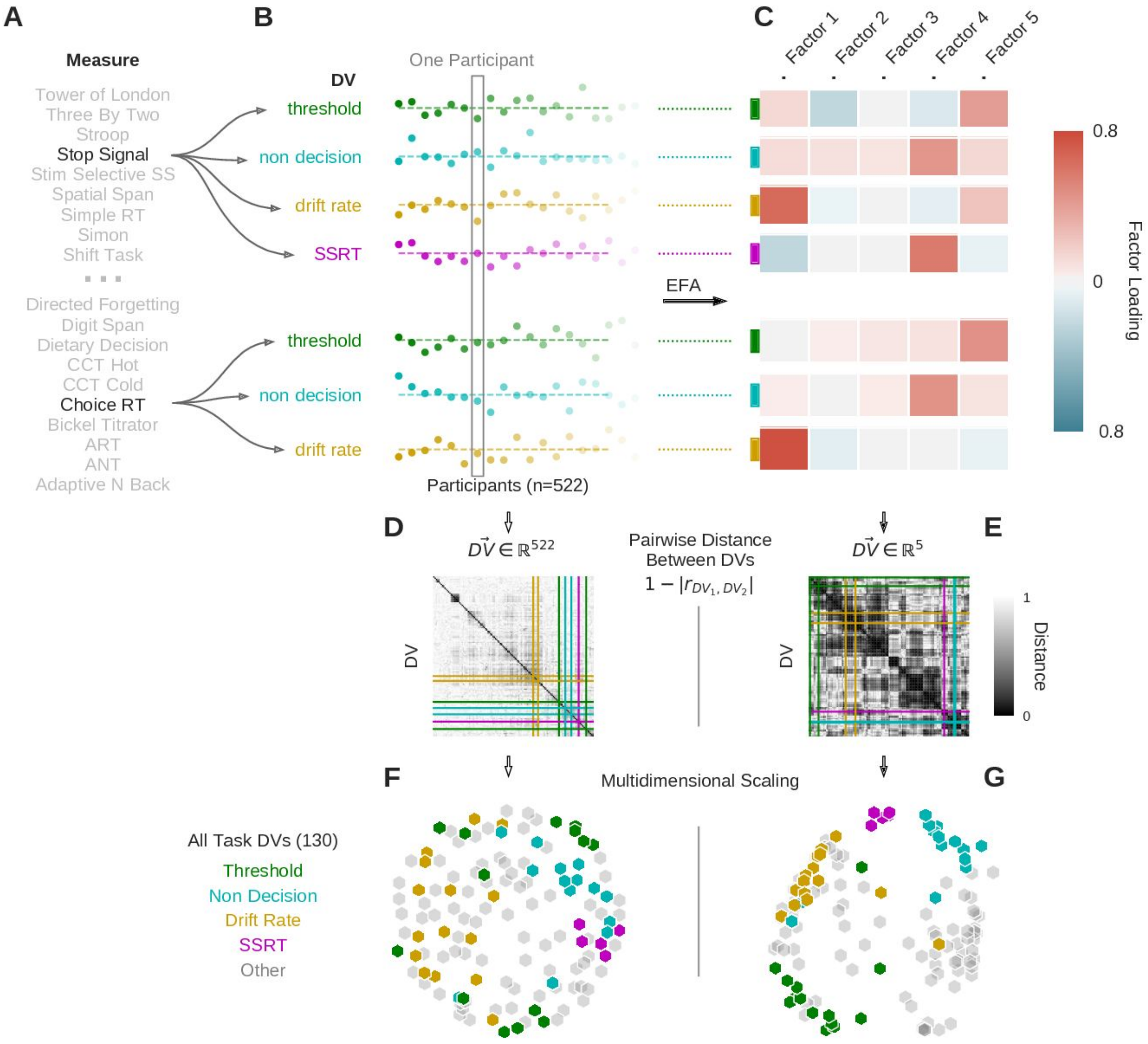


Fig. 1. Summary of Task Analytic Pipeline. (a) Participants completed 37 separate task measures, of which a subset are shown. (b) 1st-level analysis of each measure resulted in a number of DVs. Choice Reaction Time and Stop Signal are shown as two example measures, which gave rise to 7 example DVs. Participant scores are displayed as deviations from the mean for each of the 7 DVs. A subset of the 522 total participants are shown. (c) EFA projects each DV from a 522-dimensional "participant" feature space to a lower-dimensional "factor" feature space. (d-e) Pairwise-distance between all 130 task DVs are shown for the participant space (d) and factor space (e). (f-g) DV clusters are revealed in the lower-dimensional EFA space.

Multidimensional scaling of the pairwise-distances in EFA space (g) reveal obvious clustering, in contrast the the participant space (f). DVs are colored based on type for visualization purposes only - actual analysis is wholly unsupervised.

Creating a psychological space

Our first goal was to create a "psychological space": a structure that quantifies distance between DVs, and provides a vocabulary to describe disparate behavioral measurements. A foundational question is whether surveys and task DVs can be captured within a single space. Because the battery included both surveys and tasks putatively related to the same psychological constructs (e.g., impulsivity), one would predict significant relationships between the two sets of DVs, supporting a joint psychological space.

To address this goal, we evaluated the association between task and survey DVs. Neither measurement category could predict DVs from the other category, and correlations between measurement categories were weak (Figure S2; (*28*)). A psychological graph aids in visualization and demonstrates the independent clustering of the two measurement categories (Figure 2; (*28*)). The low correlations between these two groups of measures suggests a top-level ontological distinction between the constructs underlying task and survey DVs and prompted the creation of two psychological spaces.

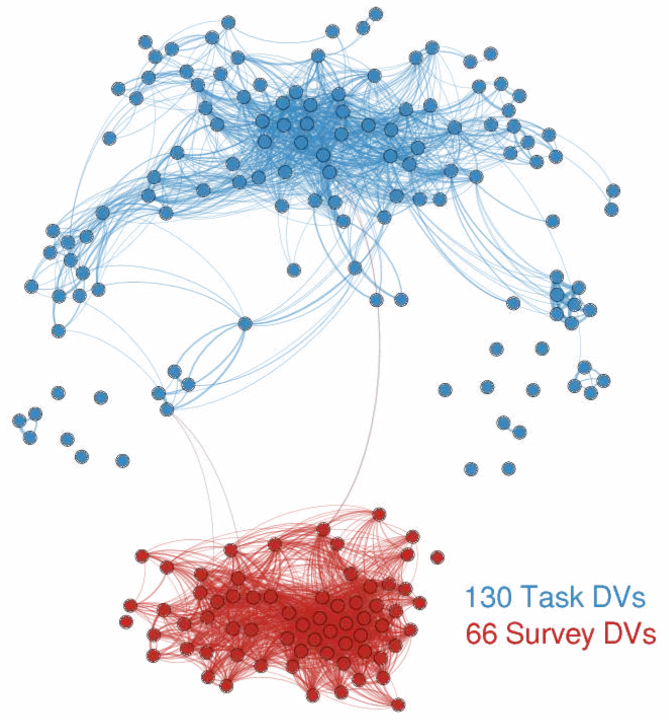


Fig. 2. Psychological graph of all DVs. Graphical lasso (*44*) was used to estimate a sparse undirected graph representing the relationships amongst all DVs. Nodes represent DVs while edges represent partial correlation between two DVs (thickness reflects strength). DVs are colored according to measurement category and edges have been thresholded (partial correlation strength >= .01).

We defined task and survey psychological spaces using exploratory factor analysis (EFA; Figure 1c; (*28*)). An important step in factor analysis is to estimate the dimensionality of the solution. Using model selection based on the Bayesian information criterion (BIC), we found that 5 and 12 factors were the optimal dimensionalities (Figure S3) for the decomposition of surveys (Figure 3) and task (Figure 4) respectively. We used oblique rotations, which revealed correlations between the discovered factors suggestive of a hierarchical factor organization (Figure S4-5). The survey EFA model fit the raw DVs better than the task EFA model (Survey $R^2 = .57$, Task $R^2 = .24$), but this difference is attenuated once test-retest reliability of individual DVs is accounted for (Figure S6; adjusted survey $R^2 = .72$, adjusted task $R^2 = .58$). Interestingly, the factor scores for both tasks and surveys demonstrated high reliability (Figure S7), which equaled (for surveys) and exceeded (for tasks) the reliability of the constituent DVs. This stability over time, a central requirement for trait measures (*29*), supports the use of factor scores as individual difference metrics.

To understand the nature of these factors we analyzed the loadings of the DVs. Briefly, a number of the survey factors (Figure S8) largely reflected separate measurement scales (e.g., Social Risk Taking and Financial Risk Taking derived from the DOSPERT) or a combination of several closely related DVs (e.g., Sensation Seeking, which related to DVs derived from the Sensation Seeking Scale, UPPS-P, I7, and DOSPERT). A notable exception was the Goal-Directedness factor, which integrates a heterogeneous set of DVs related to goal-setting, self-control, future time-perspective, and grit.

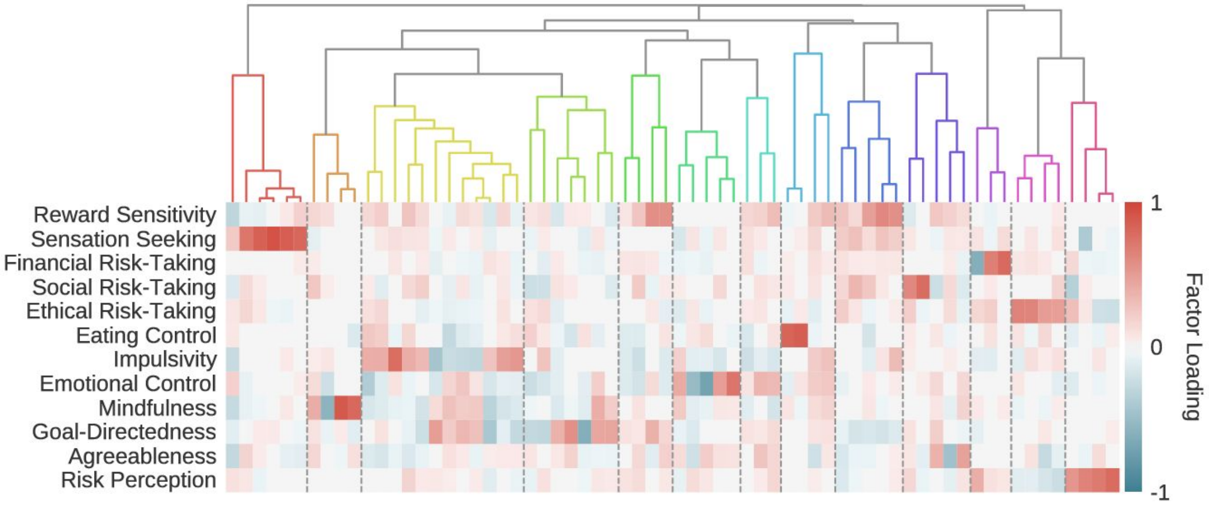


Fig. 3. Survey ontology. 66 survey DVs were projected onto 12 factors discovered using EFA, represented by the heatmap. Rows are factors and columns are separate DVs ordered based on the dendrogram above. The dendrogram was created using hierarchical clustering, and separated into clusters using DynamicTreeCut (*30*). Each cluster is separately plotted in the supplement with the DVs labeled (Figure S13).

The task EFA solution (Figure S9) was not as "simple" as the survey EFA model solution (i.e., DVs were not selectively associated with specific factors). The simplest factor was selective for temporal discounting DVs, which in turn only loaded on this factor. Three other task factors can be largely understood within the drift-diffusion model (DDM) framework; Speeded Information-Processing, Caution, and Perception/Response were strongly and differentially related to drift, threshold, and non-decision time estimates respectively. While consistent with the DDM parameterization of decision-making processes, DDM measures were not exclusively associated with these factors, as other related DV loaded sensibly (e.g., Go-NoGo d` loaded on the speeded information-processing factor). Finally, the Strategic Information-Processing factor loaded on diverse DVs that were putatively related to working-memory, general intelligence, risk-taking, introspection, and information-processing - generally tasks that were amenable to higher-order strategies, and unfolded on a time-scale greater than the speeded decision-making tasks modeled with the DDM.

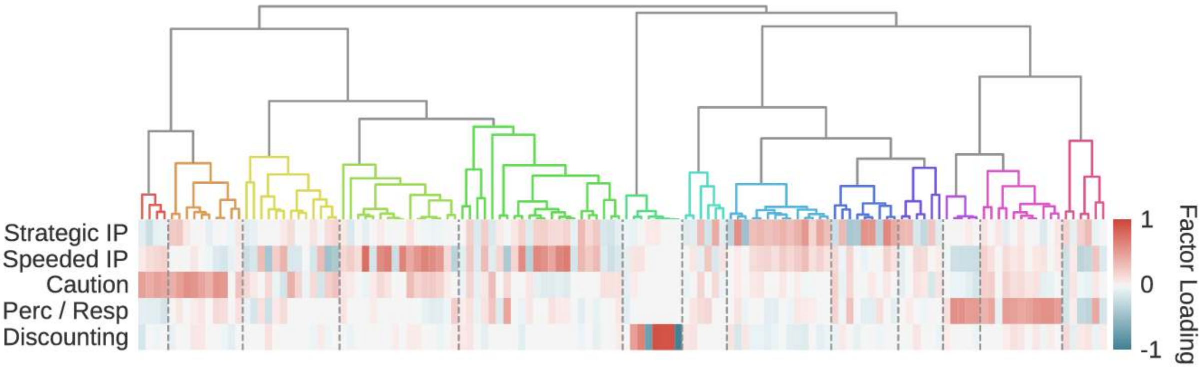


Fig. 4. Task ontology. Identical to Figure 3, except operating over 130 task DVs, which are projected onto 5 factors. Each cluster is separately plotted in the supplement with the DVs labeled (Figure S14).

Psychological Constructs as Clusters

As a whole, both task and survey factors outlined sensible psychological dimensions that relate to many concepts discussed in the field. Given this, one might ask why other plausible constructs like self-control or working memory didn't result in their own factors. However, factors should be viewed as basis vectors for a psychological space; the principal concern is the subspace *spanned* by those factors, which determines the fidelity and generalizability of the DV embedding. The particular factors are ultimately a result of rotation schemes whose goal is interpretability - a useful objective to be sure - but one potentially divorced from the span of the psychological space. A consequence is that certain psychological constructs of interest, like self-control, may emerge as clusters of DVs in this space, rather than axes. Simply put, if axes are the parametric features of a psychological space, clusters are behavioral "kinds". Both may be seen as "psychological constructs" depending on one's goals and definitions. As a result, we evaluated how DVs clustered in the spaces defined by the separate EFA models.

To identify clusters, we performed hierarchical clustering on the factor loadings of the DVs (*28*). Using this analysis, DVs that partially load on similar factors are clustered together. An alternative approach would cluster the DVs using the participant scores. We clustered DVs using the factor loadings rather than participant scores for two reasons: (1) clustering using factor loadings immediately situates each cluster within the interpretable psychological spaces defined above, and (2) projection into factor space more clearly separates DVs into meaningful and discoverable clusters (Figure 1 f,g, Figure S10-S12), suggesting that dimensionality reduction via EFA functions as a useful denoising step.

Hierarchical clustering creates a relational tree that affords clustering at multiple resolutions. To identify theory-agnostic clusters we used the Dynamic Tree Cut algorithm (*30*), which is more robust to different tree structures than simpler methods that cut the tree at one height. Doing so creates 13 clusters for both the survey DVs (Figure 3) and the task DVs (Figure 4). Of particular note in the survey clusters is the emergence of a "self-control" branch composed of two separate clusters: one primarily related to impulsivity (but also reflecting goal-directedness, mindfulness and reward sensitivity), and one reflecting long-term goal attitudes, incorporating time-perspective and implicit theories of willpower (Figure S13c,d**)**. In the task solution, a particular interesting division is between two clusters that primarily load on "strategic information processing", compared to one that loads on both "strategic" and "speeded" information processing (Figure S14h,i). See Figures S13-14 for more detail on all extracted clusters.

Prediction of Real-World Outcomes

Explicitly linking diverse psychological literatures is essential for cumulative progress in psychology, but is not sufficient. Meaningful connection to real-world outcomes is also necessary to evaluate the generalizability of psychological theories. While evaluation of "criterion validity" has historically been an important component of psychological research (though see (*7*)), ambiguity regarding outcome measures, researcher degrees of freedom, publication bias, and inadequate tests of predictive ability limit our knowledge of how psychological measures relate to real-world behaviors (*31*). The breadth of the present battery allows for a generic evaluation of the state of behavioral prediction.

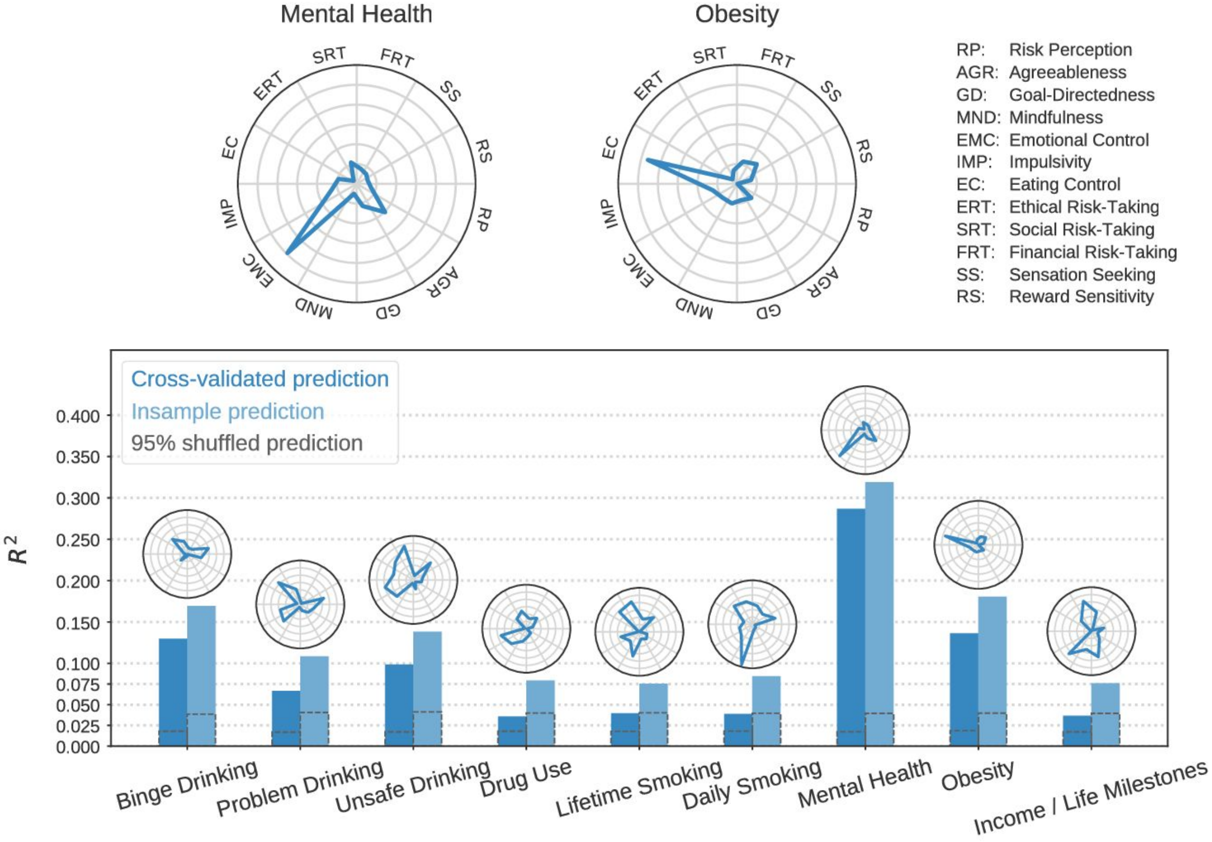
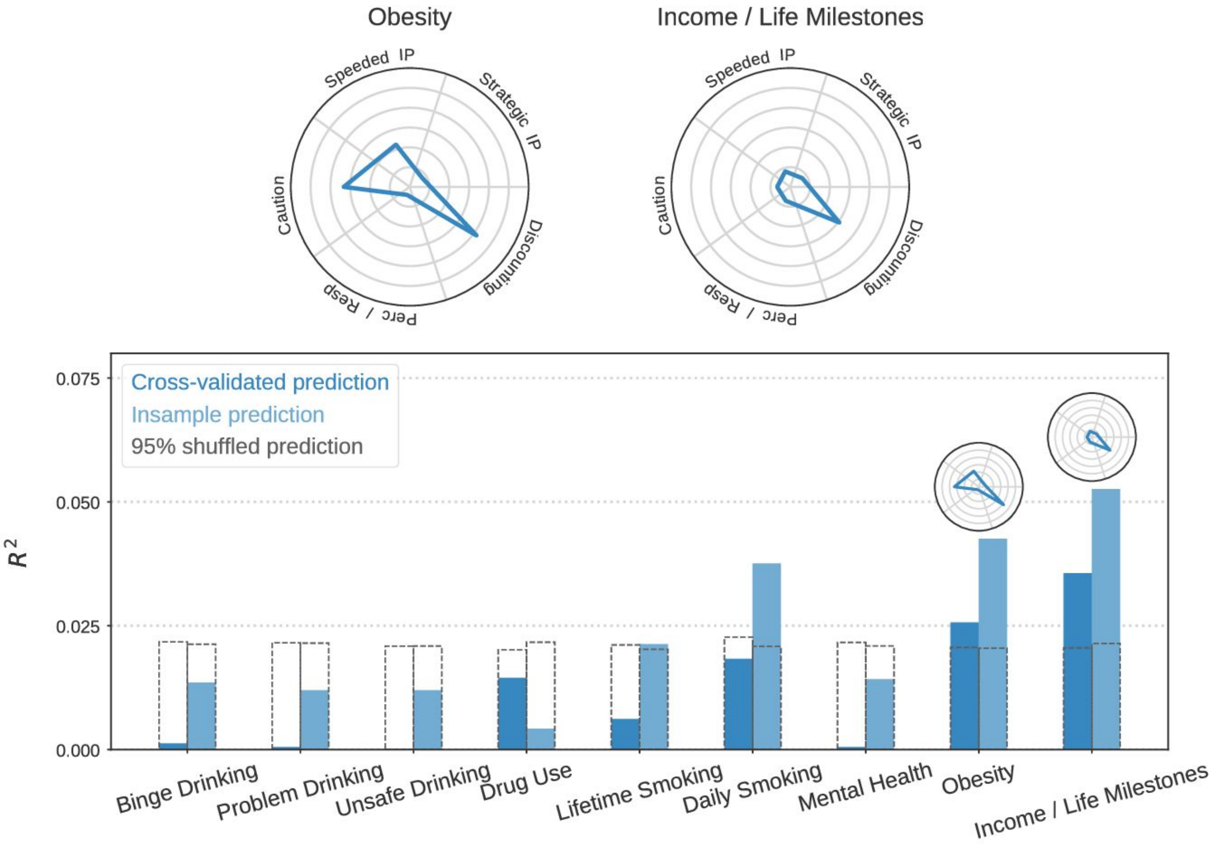


Fig. 5. Prediction of target outcomes using survey factor scores. Cross-validated (dark blue bars) and insample (light blue bars) R^2 are shown. Dashed grey boxes indicate 95% of null distribution, estimated from 2500 shuffles of the target outcome. Ontological fingerprints displayed as polar plots indicate the standardized beta value for each significant survey factor ($p < .05$). The ontological fingerprint for the two best predicted outcomes are reproduced at the top.

To evaluate predictive ability, we used a broad set of self-reported outcome measures, including socioeconomic outcomes, drug and alcohol use, and physical and mental health. We used EFA to reduce the dimensionality of the outcomes, creating "target factor scores" (referred hereafter as "targets") for each participant (Figure S15-16). Out-of-sample prediction was performed using cross validation with L2-regularized linear regression to predict targets using factor scores derived from the task and survey EFA solutions (as well as other methods (*28*), see Table S3). We created three separate predictive feature matrices: the five task factor scores, the 12 survey factor scores, and the combination of all 17 factor scores. We used factor scores rather than raw DVs due to their higher reliability, and to allow for the immediate contextualization of the predictive models within the cognitive ontology. All analyses were repeated using the raw DVs themselves as predictors, which did not change the overall interpretation (Table S4; Figure S17-18). We also performed the same analyses without cross-validation, which estimates the degree of overfitting of in-sample associations.



Obesity

Income / Life Milestones

Fig. 6. Prediction of target outcomes using task factor scores. Identical to figure 5, except for the truncated y-axis; the task factors are substantially worse at explaining variance in the target outcomes.

Surveys exhibited moderate predictive performance, significantly predicting all target variables (p<.05), with an average predictive $R^2 = .1$ (min: .05, max: .3; see Figure 5). We visualized the standardized beta coefficients of the predictive models to create an "ontological fingerprint" representing the contribution of various psychological constructs to the final predictive model for a particular target (Figures 5, 6) . The top predicted targets, mental health and obesity, have simple ontological fingerprints: "emotional control" and "problematic eating" factor scores were sufficient to predict these targets. Other fingerprints are more complicated, pointing to the contribution of multiple psychological constructs to these behaviors. The fingerprints can also be inverted, giving a sense of which kinds of real-world behaviors are related to a particular psychological construct (Figure S19).

In contrast to the surveys, tasks had almost no predictive ability (average $R^2 = .02$, max $R^2 = .04$, Figure 6). While two targets were significantly predicted above chance ($p<.05$), even for these relationships R^2 was only .04. The combined task and survey predictive model was qualitatively identical to the survey predictive model (Table S3).

Discussion

The ontological framework provides insight into mental structure by synthesizing a multifaceted behavioral dataset. Of particular note is the lack of alignment of the same putative constructs across measurement categories and low dimensionality of the psychological spaces. The former has precedent in the literature in a number of domains (*19*, *32–35*), which this work expands upon, suggesting that the inappropriate overloading of psychological terms (jingle fallacies) is widespread. Simultaneously, the low dimensionality of the psychological spaces is at odds with the varied language used by most psychologists, which obscures points of connection between literatures. Together, these findings support the need for a revised cognitive ontology.

Benefits of an ontological perspective

This approach to ontology construction fundamentally rests on correlations. Relationships amongst behavioral outputs are used to articulate mental structure, in much the same way as representational-similarity-analysis (RSA) uses neural responses to images to define the representational geometry of a brain region (*36*). To extend the analogy, much like the output of RSA serves to constrain mechanistic theories of neural function, the relationships amongst behaviors constrain more mechanistic cognitive models.

But the ontology does more than provide a large-scale "behavioral geometry" of the mind. As the ontology is a function of measurement correlations, its structure is immediately relevant for the many psychological hypotheses that are fundamentally about relationships amongst behaviors. For example, higher-order claims about the separability of various decision-making processing stages (in line with the DDM), and the discriminant validity of concepts like sensation-seeking are recapitulated by our factors. Fine-grained arbitration concerning particular measures is also possible. For instance, the angling-risk-task (ART) is widely interpreted as a risk-taking measure. However, some work suggests it may relate more to an agent's ability to assess environmental statistics and act optimally, rather than a propensity towards risky action (*34*, *37*). Our data support this latter view, as ART DVs cluster with working memory, decision-making and intelligence DVs, and are unrelated to self-report measures of risk-taking (e.g. DOSPERT).

The breadth of the dataset underlying data-driven ontology development is also important. As an example case, stop-signal reaction-time (SSRT) DVs, putatively related to response inhibition, load on the same factor as non-decision time estimates, DDM DVs intended to capture perceptual and response processes. While the clustering analysis separates SSRT from non-decision time, this suggests a relationship between these normally separable constructs. It is also apparent that without including both non-decision times and SSRT in the same measurement battery, a robust SSRT factor would be found and interpreted as “response inhibition”. Thus not only would an opportunity to bridge literatures have been missed, but a stable individual trait would have been reified.

Connecting psychological measurement to real-world behavior

The ontology also defines stable individual traits whose reliability equals or surpasses the individual DVs. For tasks in particular, EFA integrates multiple noisy DVs and creates stable measures of central psychological constructs. In doing so, EFA addresses a perennial critique of behavioral tasks: their poor psychometric properties limit their real-world applicability, particularly when it comes to predicting individual behavior (*38–40*). However, though factor scores proved reliable, they failed to predict the target outcomes, as previous work may have suggested (*41*).

Why did the surveys predict adequately, while the tasks did so poorly? The bifurcation of the ontology by measurement category suggests one explanation; tasks do not probe cognitive functions relevant for the target outcome measures. Such an explanation challenges current psychological theories of self-regulation, but allows for the possibility that the tasks would relate to other real-world outcomes. An alternative, is that the contrived nature of behavioral tasks fundamentally compromises their ecological validity (*42*). While the sensibility and reliability of the task factors speaks to real structure in human behavior, psychology's reliance on controlled experiments may lead to "theoretical overfitting". That is, theories that are explanatory and predictive of human behavior in experimental contexts may lack relevance for real human behavior. Expanding the scope of real-world outcomes would aid in distinguishing these two explanations.

In contrast, surveys predicted real-world outcomes moderately well. This may be partially explained by methodological similarity, as both surveys and the real-world outcomes in this work are self-report measures that may be susceptible to similar biases (*43*). Putting this caveat aside, the ontological fingerprints imply that these outcomes relate to an overlapping mixture of psychological constructs. If these constructs are amenable to intervention, this framework supports the development of "ontological" interventions (e.g. aimed at reducing impulsivity) that cross-cut multiple real-world behaviors. Particular behaviors like smoking could then be targeted with a multi-pronged strategy combining multiple ontological interventions. Thus the ontology holds promise for a generalizable and cumulative science of behavior change.

References and Notes

1. O. S. Collaboration, Estimating the reproducibility of psychological science. *Science*. **349**, aac4716 (2015).

2. A. Newell, You can’t play 20 questions with nature and win: Projective comments on the papers of this symposium (1973) (available at

http://repository.cmu.edu/cgi/viewcontent.cgi?article=3032&context=compsci).

3. P. E. Meehl, Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* **46**, 806 (1978).

4. P. J. Curran, The seemingly quixotic pursuit of a cumulative psychological science: introduction to the special issue. *Psychol. Methods*. **14**, 77–80 (2009).

5. J. B. L. Bard, S. Y. Rhee, Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* **5**, 213–222 (2004).

6. R. A. Poldrack, T. Yarkoni, From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annu. Rev. Psychol.* **67**, 587–612 (2016).

7. D. Borsboom, G. J. Mellenbergh, J. van Heerden, The concept of validity. *Psychol. Rev.* **111**, 1061–1071 (2004).

8. L. J. Cronbach, P. E. Meehl, Construct validity in psychological tests. *Psychol. Bull.* **52**, 281–302 (1955).

9. T. Insel *et al.*, Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry*. **167**, 748–751 (2010).

10. T. Yarkoni, J. Westfall, Choosing prediction over explanation in psychology: Lessons from machine learning. *FigShare, https://dx. doi. org/10. 6084/m9. figshare*. **2441878**, v1 (2016).

11. T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning. 2001. *NY Springer* (2001).

12. J. B. Copas, Regression, prediction and shrinkage. *J. R. Stat. Soc. Series B Stat. Methodol.*, 311–354 (1983).

13. D. W. Russell, In Search of Underlying Dimensions: The Use (and Abuse) of Factor Analysis in Personality and Social Psychology Bulletin. *Pers. Soc. Psychol. Bull.* **28**, 1629–1646 (2002).

14. J. M. Digman, Personality Structure: Emergence of the Five-Factor Model. *Annu. Rev. Psychol.* **41**, 417–440 (1990).

15. A. S. Cowen, D. Keltner, Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7900–E7909 (2017).

16. S. R. Briggs, J. M. Cheek, The role of factor analysis in the development and evaluation of personality scales. *J. Pers.* **54**, 106–148 (1986).

17. B. D. Haig, Exploratory Factor Analysis, Theory Generation, and Scientific Method. *Multivariate Behav. Res.* **40**, 303–329 (2005).

18. A. Miyake *et al.*, The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: a latent variable analysis. *Cogn. Psychol.* **41**, 49–100 (2000).

19. C. Stahl *et al.*, Behavioral components of impulsivity. *J. Exp. Psychol. Gen.* **143**, 850–886 (2014).

20. H. P. Kotabe, W. Hofmann, On Integrating the Components of Self-Control. *Perspect. Psychol. Sci.* **10**, 618–638 (2015).

21. T. E. Moffitt *et al.*, A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2693–2698 (2011).

22. I. W. Eisenberg *et al.*, Applying novel technologies and methods to inform the ontology of self-regulation. *Behav. Res. Ther.* **101**, 46–57 (2018).

23. A. L. Duckworth, M. L. Kern, A Meta-Analysis of the Convergent Validity of Self-Control Measures. *J. Res. Pers.* **45**, 259–268 (2011).

24. D. W. Schneider, G. D. Logan, Task-switching performance with 1:1 and 2:1 cue-task mappings: not so different after all. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 405–415 (2011).

25. U. Mayr, R. Kliegl, Differential effects of cue changes and task changes on task-set selection costs. *J. Exp. Psychol. Learn. Mem. Cogn.* **29**, 362–372 (2003).

26. T. V. Wiecki, I. Sofer, M. J. Frank, HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front. Neuroinform.* **7**, 14 (2013).

27. V. V. Sochat *et al.*, The experiment factory: Standardizing behavioral experiments. *Front. Psychol.* **7** (2016), doi:10.3389/fpsyg.2016.00610.

28. Materials and methods are available as supplementary materials at the Science website.

29. E. L. Hamaker, J. R. Nesselroade, P. C. M. Molenaar, The integrated trait–state model. *J. Res. Pers.* **41**, 295–315 (2007).

30. P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*. **24**, 719–720 (2008).

31. S. Michie, M. Johnston, Theories and techniques of behaviour change: Developing a cumulative science of behaviour change (2012) (available at http://www.tandfonline.com/doi/full/10.1080/17437199.2012.654964).

32. M. A. Cyders, A. Coskunpinar, Measurement of constructs using self-report and behavioral lab tasks: is there overlap in nomothetic span and construct representation for impulsivity? *Clin. Psychol. Rev.* **31**, 965–982 (2011).

33. R. K. McHugh *et al.*, Shared Variance among Self-Report and Behavioral Measures of Distress Intolerance. *Cognit. Ther. Res.* **35**, 266–275 (2011).

34. R. Frey, A. Pedroni, R. Mata, J. Rieskamp, R. Hertwig, Risk preference shares the psychometric structure of major psychological traits. *Sci Adv*. **3**, e1701381 (2017).

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

36. N. Kriegeskorte, R. A. Kievit, Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).

37. J. D. Jentsch, J. A. Woods, S. M. Groman, E. Seu, Behavioral characteristics and neural mechanisms mediating performance in a rodent version of the Balloon Analog Risk Task. *Neuropsychopharmacology*. **35**, 1797–1806 (2010).

38. A. Z. Enkavi *et al.*, A large scale analysis of test-retest reliabilities of self-regulation measures. *Manuscript In Preparation* (2018).

39. C. Hedge, G. Powell, P. Sumner, The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* (2017), doi:10.3758/s13428-017-0935-1.

40. S. O. Lilienfeld, M. T. Treadway, Clashing Diagnostic Approaches: DSM-ICD Versus RDoC. *Annu. Rev. Clin. Psychol.* **12**, 435–463 (2016).

41. L. Sharma, K. E. Markon, L. A. Clark, Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychol. Bull.* **140**, 374–408 (2013).

42. P. Anderson, Assessment and development of executive function (EF) during childhood. *Child Neuropsychol.* **8**, 71–82 (2002).

43. A. Furnham, M. Henderson, The good, the bad and the mad: Response bias in self-report measures. *Pers. Individ. Dif.* **3**, 311–320 (1982).

44. J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. **9**, 432–441 (2008).

Acknowledgments:

We thank V. Sochat for developing the experiment factory infrastructure, and J. Wright, K.J. Gorgolewski and the SOBC network for discussions and suggestions. **Funding**: 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. **Author contributions:** Conceptualization: I.W.E., P.G.B., A.Z.E, D.P.M., L.A.M., R.A.P.; Methodology: I.W.E., P.G.B., A.Z.E., D.P.M, R.A.P.; Software: I.W.E., A.Z.E., J.L.; Formal Analysis: I.W.E., A.Z.E., R.A.P.; Investigation: I.W.E., A.Z.E., J.L.; Resources: R.A.P.; Data Curation: I.W.E., P.G.B.; Writing: I.W.E., P.G.B., R.A.P., Visualization: I.W.E., Supervision: P.G.B., R.A.P.; Funding Acquisition: L.A.M., R.A.P, **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All data is available in the github repository related to this work: https://goo.gl/uzgQUZ

Supplemental Materials

Materials and Methods

Many of the methods in this project have been previously documented in a protocol paper summarizing our research program focused on behavior change (*1*). For convenience, we have reused text from that paper in this supplement.

Experimental Design

Data Acquisition and Extraction of Individual Difference Measures

To begin the enterprise of data-driven ontology development, we collected a dataset of 522 adult participants, each completing a battery composed of 37 behavioral tasks and 23 self-report surveys. As such, it included measures putatively related to self-regulation including risk-taking, temporal discounting and impulsivity, but also extended into more generic cognitive domains like working memory, information processing, learning, mindfulness, and others. By construction, some putative constructs like impulsivity were evaluated in both surveys and tasks, affording the opportunity to evaluate cross-measure consistency. In addition to these surveys and tasks, participants reported a number of "real-world" outcomes (e.g. questions relating to alcohol consumption, mental health, personal finances, etc.). Data was collected on Amazon Mechanical Turk using the Experiment Factory platform (*2*). The Experiment Factory allows for easy replication or extension of this dataset, and the current dataset is freely available (https://goo.gl/uzgQUZ). The data acquisition plan was pre-registered on the Open Science Framework (http://goo.gl/3eJuu1)

Mechanical Turk Data Collection Procedure

The dataset used in this analysis was collected as part of a larger project investigating self-regulation and behavioral change, outlined in our previous protocol paper (*1*). Our analysis plan originally was divided into a discovery (N=200) and validation (N=300) cohort. Though we collected the dataset in this format, and constructed most of the analytic pipeline based on the discovery data, our analyses presented here are on the entire dataset. Specifically, the analysis plan for the majority of the behavioral and self-report measures (e.g. the selection and operationalization of the various dependent variables, specification of quality control) was decided on prior to unblinding the validation dataset. After unblinding, some dependent variables were changed, either due to the discovery of coding errors, or the recognition that we had missed canonical analyses for individual measures. Importantly, our DV calculations were never informed by the subsequent structure-discovery analyses. Following collection of the discovery and validation cohorts, we retested a subset of participants (n=150). The retest subset was selected randomly from the discovery and validation cohort and completed the battery a second time. The battery required roughly 10 hours to complete.

Due to the large number of participants, and significant time involvement, we used Amazon's Mechanical Turk (MTurk) to collect behavioral data. In contrast to most psychological studies on MTurk, which consist of a single relatively short testing session, the behavioral battery required multiple sessions due to its length. To address this issue, we developed the Experiment Factory (*3*), an infrastructure to deploy behavioral measurements on MTurk. The Experiment Factory presented tasks to participants in a random order, and allowed participants to complete the battery at their own pace, finishing as many or as few tasks as they wanted in each sitting. Participants were required to finish the entire battery within one week of accepting the HIT, but no other restriction was placed on their time. Only adults who had completed 2000 previous HITs with a 95% approval, and were between 18-50 and living in the US were invited to participate, though four participants reported that their age was between 50-60. For completion of the battery, participants were paid $60 plus bonuses from performance on specific tasks averaging $10 for their time (minimum: $65, maximum: $75).

As the behavioral battery was long (both in comparison to other psychology studies and MTurk HITs), reducing attrition was a significant consideration. In order to minimize attrition, a number of steps were taken, including providing comprehensive instructions, follow-up emails, and actively fielding questions on various online message boards for MTurk workers. Also, as an incentive to complete, we created a payment schedule that paid a lower rate if the participant failed to complete all 63 measures in the battery. Together, these steps kept attrition manageable: 84**%** of all participants who enrolled ultimately completed the entire battery. We removed any participants who failed to complete the entire battery (102 out of 662), as well as any who failed to pass quality checks (see "Quality Checks for Cognitive Tasks", below) and continued recruiting until we achieved our sample size goal for each cohort. Due to over-recruiting to ensure we achieved minimum sample sizes, our final samples were 200 (discovery) and 322 (validation; all extra participants beyond the planned sample were assigned to the validation set). Finally, completed participants were iteratively solicited to take the entire battery a second time, until 150 completed the battery while passing quality checks. Completed participants were randomly ordered before solicitation, and all participants completed the retest within 8 months of the initial test (minimum 60 days, maximum 228 days gap between completions)."

Quality Checks for Cognitive Tasks

Participants on MTurk are wholly unsupervised, necessitating procedures to ensure data quality. Quality checks were broadly applied to all cognitive tasks to ensure that (1) response times were not unreasonably fast on average, (2) omitted responses were reasonably low, (3) accuracy on cognitive tasks was reasonably high and (4) responses were sufficiently distributed (i.e. the participant didn't only press a single key). The specific criteria we used differed for some tasks, but in general we required that median response times were longer than 200 ms, no more than 25% of responses were omitted, accuracy was higher than 60% and no single response was given more than 95% of the time. These thresholds were set based on evaluation using the discovery cohort only, prior to unblinding the validation cohort. Overall, these steps were taken to ensure that participants in our dataset completed the tasks in earnest. Similar checks could not be performed on the self-report surveys or demographic measurements as we did not collect response time measures and potentially suspect response patterns (e.g. selecting only one response for every item) may be input honestly.

These criteria were used to evaluate each participant/task pair; failure on any check led to removal of that particular task's data for that participant. In addition, we removed a participant's entire dataset if they failed on four or more individual tasks (38 out of 560 participants were so removed).

These quality checks were intended as thresholds to screen out participants who were intentionally gaming the HIT. We also used task-specific manipulation checks which evaluated particular performance criteria specific to different tasks, necessary for the interpretability of our derived dependent measures. Failing these manipulation checks led to the removal of that participant's data on the failed task, but did not count towards the four failed tasks that would lead to the entire participant being removed from our study. The tasks that used these additional manipulation checks were the stop signal tasks, probabilistic selection task, and two-step decision task.

Selection of Dependent Variables

From the 37 tasks and 23 surveys we computed 205 dependent variables (DVs). Each survey was analyzed identically - canonical subscale scores were used as DVs. That is, items were appropriately scored (and reversed, if necessary) and summed or averaged in accordance with individual survey scoring procedures.

The tasks were heterogeneous, preventing a completely generic analysis strategy. Nonetheless, many tasks involved speeded decisions between two alternatives, and are well characterized by reaction time and accuracy. It is well known that reaction time and accuracy are confounded by the speed-accuracy tradeoff (*4*), which prompted us to use the drift-diffusion model (DDM). The basic DDM transforms accuracy and reaction time into a drift rate, threshold, and non-decision time, roughly corresponding to performance, response caution (a point along the speed-accuracy tradeoff curve) and stimulus-processing/motor-planning, respectively. We fit the DDM parameters using the hierarchical drift-diffusion model (HDDM). HDDM models the DDM parameters hierarchically, such that individual parameters are assumed to be drawn from a group distribution (*5*). This procedure improves data efficiency (*6*), and has been shown to better capture true parameters when dealing with small datasets, or datasets corrupted by trials influenced by processes other than evidence accumulation (e.g., attentional lapses). Though individual parameter estimates are no longer independent (due to the hierarchy), hierarchical models also have been shown to improve point estimates of individual parameters, and are particularly useful when one is interested in correlations between other traits and the individual parameter estimates (*7*). The HDDM also allows DDM parameters to modeled as a function of various conditions. For example, when modeling the stroop task, we modeled drift rate as a function of conflict condition while keeping the other parameters constant.

Tasks that were not speeded choice tasks were heterogeneous and each analyzed according to its own scientific tradition. The full list of measures is available in Table S1 (surveys) and Table S2 (tasks).

Data cleaning and imputation

To ensure we did not have redundant variables in the participant-by-measure data matrix, if any two dependent variables derived from the same task or survey measure were correlated $r > 0.85$, one of the variables was arbitrarily removed. 6 variables were dropped using this criteria. In addition, because many of our analyses assume normally distributed variables, we transformed skewed variables (absolute skew > 1) and removed any variable that remained excessively skewed after transformation. 3 variables were dropped due to non-normality. Finally, our data matrix had missing values due to our quality check procedure. Only 3.1% of the overall data matrix was missing, but these missing values were not uniformly distributed amongst the DVs. Instead, 47% of the DVs had no missing values, while a small subsection (the stop signal tasks, probabilistic selection task and two-step decision task) had substantially more missing values (between 10%-30%) due to the additional quality control measures (manipulation checks) taken on those particular tasks. We imputed the data matrix using R's missForest package (*8*). See Table S1 and Table S2 for specification of which variables were transformed or dropped due to these procedures.

Statistical Analysis

Assessment of Test-Retest Reliability

The battery was notably divided by measurement type; self-report surveys all required the participant to answer one or more questions as honestly as they could with no time pressure, while the tasks varied considerably in their design, but generally required sustained attention and had an objective notion of performance. More importantly, however, was the potential differences in psychometric properties between these measurement types. Surveys are generally developed with psychometric theory in mind, and are routinely assessed for reliability. This is in stark contrast with tasks, where psychometric properties are often unknown, and rarely reevaluated.

We evaluated reliability of all DVs in the same population by analyzing the subset of participants that completed the entire battery a second time. The full description of the subsequent analysis are laid out in Enkavi et al (*9*); summarizing, the surveys showed greater test-retest reliability (ICC $M = .87$, $SD = .06$) compared to the tasks (ICC $M = .52$, $SD = .21$), see Figure S1. There was substantial heterogeneity within tasks with some measures (e.g. discounting and DDM parameters) performing much more reliably than others. We used Pearson correlations between the two sessions as a measure of reliability, establishing a "noise-ceiling" (maximum predictive power possible given irreducible noise) to evaluate the fit of the exploratory factor analyses.

Association Between Tasks and Surveys

Task and survey DVs had weak to no relationship with each other, as is evident by their uncorrected Pearson correlations (Figure S2a). To more rigorously quantify the relationship between tasks and surveys we employed two separate methods. First, we assessed how well a held out DV was predicted by either all task or survey DVs (excluding the to-be-predicted DV). This resulted in 4 distributions of predictions: two within-measurement predictions (task-by-tasks and survey-by-surveys) and two across-measurement predictions (task-by-surveys and survey-by-tasks).

Prediction success was assessed by 10-fold cross-validated ridge regression using the RidgeCV function from scikit-learn with default parameters (*10*). Almost all DVs were able to be predicted to some degree by their respective measurement category (mean task-by-tasks $R^2 = .22$, mean survey-by-surveys $R^2 = .45$). In contrast, cross-measurement prediction failed: surveys were unable to predict task DVs and vice versa (mean task-by-surveys $R^2 = -.13$, mean survey-by-tasks $R^2 = -.29$). Note that R^2 values below 0 are possible when employing cross-validation and indicate no discoverable linear relationship. The entire distribution of R^2 values for each of these predictions is shown is Figure S2b.

We also assessed this relationship by constructing a psychological graph, where nodes are unique DVs and edges reflect the partial correlation between two DVs after conditioning on all other DVs. To estimate these correlations, we employed the Graphical Lasso (*11*) using the EBICglasso function from the QGraph package (*12*). Visualization of the graph (Figure 2) was accomplished using a force-directed algorithm in Gephi (*13*), with edges reflecting the absolute value of the partial correlations greater than .01. While edge strength varies between different pairs of DVs within the same measurement category, reflective of complex psychological structure, all edges between measurement categories are close to 0 (Figure S2c).

Exploratory Factor Analysis

Exploratory factor analysis (EFA) seeks to explain the covariability of a number of observed variables in terms of a smaller number of latent (unobserved) variables, called factors. Each observed variable is modeled as a linear combination of these latent factors and some measurement error:

$$x - \mu = LF + \varepsilon$$

where x is an M (DV) x N (participant) matrix of observed DVs, μ is a matrix of variable means, L is the M x f (*number of factors*) loading matrix, describing the relationship between each variable and the latent factors, and F is the f x N matrix of factor scores. ε captures measurement error - the variance left unexplained by the latent (common) factors. In the current study each DV is represented by 522 participants - the individual participant scores - and EFA is used to estimate the embedding of these DVs (represented by the loading matrix) in a common psychological space spanned by the latent factors. Once estimated, factor scores are computed, representing the degree to which an individual represents that latent factor. For example, we used EFA to reduce the outcome measures to 9 factors, which are then used to compute 9 factor scores for each participant. These factor scores became the "outcome targets". One outcome target heavily loaded variables related to binge drinking, and did not highly load any other variable - thus its related factor score represents an individual's general tendency to binge drink, and was named accordingly.

EFA was performed using maximum likelihood estimation, followed by oblimin rotation to rotate the factors without enforcing orthogonality. Factor rotation leads to easier interpretation by optimizing "very simple structure" (*14*), without changing the fit of the model. Factor scores were estimated using the "tenBerge" method, which is most appropriate given oblique rotation (*15*). All analyses were implemented using the "fa" function from the psych package in R (*16*).

An important step when performing EFA is deciding on *f*, the number of factors to estimate. While there are many procedures to accomplish this, we chose the number of factors that minimized the Bayesian Information Criteria (BIC). BIC is a criterion for model selection that attempts to correct for overfitting by penalizing more complex models, with lower values represented a better balance between capturing the data and model complexity. Figure S3 displays BIC values for EFA solutions with different numbers of factors. Other criteria identified an overlapping range of optimal dimensionalities, consistent with the notion that there is no single "best" dimensionality (*17*).

The optimal solutions for tasks, surveys and outcome variables all had significant cross-factor correlations after oblimin rotation. This observation prompted us to investigate hierarchical factor solutions - fitting a 2nd-level EFA model on top of the 1st-level factors (Figures S6-8). This 2nd-level factor analysis does not affect the fit to the original data in any way - instead it merely serves as an interpretive tool to understand the correlation across factors. Estimation of the number of higher-order factors and the 2nd-level factor model proceeded identically to the 1st-level EFA.

Factor Analysis Communality and DV Test-Retest Reliability

Communality refers to the variance accounted for in the DVs by the EFA model. Average communality (equivalent to overall variance explained by the EFA model) was greater for survey DVs (*M* = .57, *SD* = .18) than task DVs (*M* =.24, *SD* = .19), and differed between different DVs. Though this is partially explained by the different number of factors identified using the BIC criterion (5 factors for tasks, 12 factors for surveys), a 12 factor task model still only had an average communality of .32.

There are two main explanations for low communality: either the estimated factors do not span a psychological space that properly represents all DVs (e.g., the factors are a poor model for the data) or the DVs themselves have poor measurement characteristics. The latter creates a "noise ceiling", and puts an upper bound on the variance that can be explained by any model

To investigate this we correlated communality and test-retest reliability (as measured by Pearson correlation). We only evaluated DVs which had a test-retest reliability above .2. We found a strong correlation between communality and test-retest reliability in the tasks (r = .62), and a smaller correlation between communality and test-retest reliability in surveys (r = .39), suggesting that measurement characteristics are related to differential communality across DVs. We adjusted for test-retest reliability by dividing the communality values for individual DVs by their test-retest reliability, which results in an "adjusted" measure of variance explained. After adjustment the task factor model explained 58% of the explainable variance (across DVs *SD* = .32) , while the survey factor model explained 72% (across DVs *SD* = .21) (Figure S6). Thus the discrepancy in explained variance can largely be understood in terms of the poor measurement properties of task DVs.

Factor Score Reliability

Though task DVs were less reliable than surveys in general, it was possible that factor scores derived from the task EFA model were just as reliable as the survey factor scores. The intuition is that by integrating over many noisy measurements of a central psychological construct, EFA creates a stable individual trait, much as survey summary scores are more reliable than the specific items that constitute that scale.

To evaluate this we made use of the 150 participants who completed the entire battery a second time (see Assessment of test-retest reliability). Factor scores were computed at both time points making use of the weight matrix derived from EFA run on the first completion (i.e. the same linear combination of DVs was used to create factor scores at both time points). Reliability was quantified by the Pearson correlation between factor scores at both time points. All 5 task factors (*M* = .81, min = .76, max = .86) and 12 survey factors (*M* = .86, min = .75, max = .95) proved highly reliable (Figure S7).

Hierarchical Clustering

Hierarchical clustering is a family of algorithms that builds a relational tree. We used an agglomerative clustering technique that iteratively combines DVs (separately for surveys and tasks). This technique relies on a predefined distance metric which defines how clusters should be combined. Inspired by representational similarity analysis (*18*), we used correlation distance (or dissimilarity) as our distance metric. Because of the arbitrary direction of our measures (e.g. an "impulsivity" DV could easily be represented by a flipped "self-control" DV) we used absolute correlation distance, defined as:

$$distance = 1 - |r|$$

We did not compute the correlation distance in native (participant) space, but rather in the factor analytic embedding space defined by the loading matrix. That said, following this initial analysis, we performed hierarchical clustering in native space for both surveys and tasks, and the dendrograms are shown in Figure S10 and Figure S11, respectively. It is clear from the dendrograms that the clustering is worse in this space, which is reflected analytically using silhouette analysis in Figure S12b,d.

The hierarchy created by this technique has no intrinsic cut points, and thus no objective clusters. To identify clusters which are interpretively useful, we used the "dynamic hybrid cut algorithm" from the DynamicTreeCut package (*19*). In comparison to naive approaches, which cut the dendrogram at a particular height to identify clusters, the dynamic tree cut algorithm cuts the tree at different heights depending on the structure of the underlying branch. We separately evaluated clustering using a simpler partitioning algorithm - cutting the tree at a single height in order to maximize the mean silhouette score. At most heights the silhouette score is comparable to the dynamicTreeCut clustering solution, except at very low cut heights, which produce many small, uninterpretable clusters (Figure S12). Finally, if we compare the clustering solution produced by dynamicTreeCut to a single height cut that produces the same number of clusters we find good convergence between the clustering solutions, as quantified by the adjusted mutual information score (AMI) between the two clustering solutions (task *AMI* = .91; survey *AMI* = .88).

Prediction Analysis

The primary prediction analysis used the factor scores from tasks or surveys, as well as both combined, as features to predict outcome targets. These outcome targets were derived from EFA on the individual outcome items (e.g. household income, cigarette habits) which yielded 9 factors: Binge Drinking, Problem Drinking, Unsafe Drinking, Drug Use, Lifetime Smoking, Daily Smoking, Obesity, Mental Health, and Income/Life Milestones. We used two different regularized linear regression methods to perform prediction: lasso and ridge regression, which differ in the form of their regularization. We also used two nonlinear regression methods: random forest and support vector machines. All methods used scikit-learn (*10*).

Cross-validation was performed using a balanced 10-fold procedure (custom code based on (*20*)), thus fitting each model with 469 participants and testing on 53 left out participants. Across all folds each participant's demographic factor scores (the prediction targets) were predicted in a cross-validated manner. These estimates were correlated with the actual demographic factor scores to compute R^2. Insample R^2 were estimated by fitting identical models as above to the whole dataset and testing on the same dataset. Mean absolute error (MAE) was computed analogously. Cross-validated and insample R^2 and MAE for each model are shown in Table S3. Ridge and lasso regression performed comparably, while nonlinear methods, particular random forests, overfit the data producing poor fits. Ridge regression was used to assess feature importance due to its desirable regularization properties compared to lasso (sparse feature selection was not necessary for interpretability with so few predictors) and comparable performance. "Feature importance" for the ridge regression (as shown in the "ontological fingerprint" polar plots in Figure 5, 6, S19) is defined as the standardized beta coefficients.

Prediction results combining task and survey factor scores did not differ qualitatively from the prediction results using survey factor scores alone, except for a slight improvement for obesity and income/life-outcomes, which were the two targets where tasks performed above chance (Table S3).This constitutes weak evidence that, for some targets, tasks can complement surveys to create a predictive model for real-world behavior.

One potential issue with our prediction analysis is the possibility of data-bleeding between cross-validation folds as a result of the factor analytic models. That is, the EFA models for both predictors (e.g. survey factor scores) and targets (outcome target factor scores) were fit on the entire dataset. This data-bleeding could inappropriately inflate prediction estimates. To control for this possibility we created an empirical "null" distribution of prediction success by shuffling the target outcomes and repeating the prediction 2,500 times. 95% prediction success is shown in all prediction plots and is used as a significance cut off ($p < .05$) to display ontological fingerprints.

Complementing our prediction using task and survey factor scores derived from EFA, we performed the same analyses using the individual DVs (separately for tasks and surveys) as predictor features. Cross-validated results are qualitatively the same as the EFA analysis using ridge regression. Lasso also showed qualitatively similar results (Figure S17-18), though quantitatively differed on specific targets. In particular, the target factor "Binge Drinking" was better predicted by Lasso with the survey DVs ($R^2 = .25$) than using the survey factor scores ($R^2 = .13$). Due to the variable selection imposed by lasso, only four DVs contributed to this prediction (TFEQ-R18: Cognitive Restraint, SSS: Disinhibition, ZTPI: Past Positive, DOSPERT: Healthy Safety Risk-Taking). This demonstrates a difficulty inherent in building predictive models using individual DVs - it is difficult to know how to generalize predictive success or how to connect prediction to theoretical constructs. While this again highlights the utility in making use of ontological factors, it also demonstrates that if prediction is the only goal, there are times when dimensionality reduction is deleterious. The full prediction results for both linear models using the DVs as predictors is shown in Table S4. Overall, the qualitative agreement between prediction results with or without using EFA indicates that EFA did not generally remove information pertinent for outcome prediction.

Description of Self-Report Surveys

The description of the individual measures borrows text from Enkavi et al. (*9*). Many of the measures have also been described on the Science of Behavior Change's website, and can be demoed there. In addition, the specific items and coding can be found at the expfactory-survey page, and the survey subscale scoring (the particular items used for each subscale) can be found within the expfactory-analysis repo. Data on individual surveys can be found in the Self Regulation Ontology repo.

Behavioral Inhibition and Approach (BIS/BAS)

Developed by Carver and White (*21*) to measure behavioral approach and inhibition systems, BIS/BAS is a 24 item scale 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’). Questions are presented with four point scales.

Barratt Impulsiveness Scale (BIS-11)

BIS-11 (*22*) is a 30 item questionnaire using a four point scale for short questions. Factor analyses reveal six first order factors that can be further grouped into three second order factors. The first order factors are attention (‘I “squirm” at plays or lectures’) and cognitive stability (‘I often have extraneous thoughts when thinking’) that constitute the second order attentional factor, motor (‘I act “on impulse”’) and perseverance (‘I change residences’) that constitute the motor second order factor and the self-control (‘I am a careful thinker’) and cognitive complexity (‘I like to think about complex problems’) factors that constitute the nonplanning second order factor.

Brief Self-Control scale (BSCS)

BCSC (*23*) is a 13 item scale presented with 5 point scales (1: Not at all to 5: Very much) that measures self-control. An example item is "I am good at resisting temptation."

Dickman's Functional and Dysfunctional Impulsivity

This survey (*24*) distinguishes between two types of tendencies to act without forethought: one that has negative consequences (dysfunctional) and one that is more optimal (functional). The dysfunctional impulsivity factor consists of 12 true/false items (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’) and the functional impulsivity factor consists of 11 true/false items (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’).

Domain specific risk taking (DOSPERT - RT/RP/EB)

DOSPERT (Domain Specific Risk Taking) survey attempts to capture a more comprehensive, interpretable and translatable construct of risk attitude that is reduced to a single number across domains and confounds marginal value for outcomes and attitudes towards risk in frameworks based on the expected utility theory. The abbreviated version (*25*) consists of 30 scenarios that are presented with slight variations in question wording to form three separate subscales intended to detangle these. In the risk taking (RT) subscale participants are asked the likelihood they would engage in the described activity; in the risk perception (RP) subscale they are asked how risky they assess each situation to be and finally in the expected benefits (EB) 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.'), social (S; "Admitting that your tastes are different from those of a friend.'). All items are presented with a 7 point scale.

Emotion Regulation Questionnaire (ERQ)

Developed by Gross and John (*26*) 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.

Five Facet Mindfulness Questionnaire (FFMQ)

FFMQ is a result of a broad psychometric analysis of multiple mindfulness questionnaire. Baer et al. (*27*) 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.'), nonjudging 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.

Future Time Perspective (FTP)

Developed by Carstensen and Lang (*28*) in the context of socioemotional selectivity theory and related to the SOC questionnaire 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.

Grit Scale (GRIT-S)

Developed by Duckworth and Quinn (*29*) 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'). We used the total score as a single "grit" DV.

I-7 impulsiveness and venturesomeness questionnaire

The culmination of Eysenck’s work in developing an impulsivity questionnaire the I-7 is the most recent version following I-5 and I-6 (*30*). Though the scale 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.

Mindful Attention and Awareness Scale (MAAS)

Developed by Brown and Ryan (*31*) 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.

Multidimensional Personality Questionnaire (MPQ) Control Scale

The MPQ (*32*) is a comprehensive and long questionnaire consisting of multiple subscales. We only used the 24-item single factor control subscale adopting the strategy of Whiteside and Lynam (*33*). Typical true/false items for the MPQ are "I am fast and careless." or "I do things on the spur of the moment.

Zuckerman's Sensation Seeking Scale (SSS-V)

This scale (*34*) is intended to measure the concept of optimal stimulation level. Participants are presented with two scenarios in each question and asked to indicate which they would prefer. Zuckerman (*35*) identified four factors that the scale 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’), 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.

Selection-Optimization-Compensation (SOC) questionnaire

This questionnaire is developed as a measurement tool of a metatheory of life management strategy within lifespan psychology. Developed by Baltes et al. (*36*) it is intended to measure four components: 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’) that together constitute the selection component, 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’). Each item presents two scenarios that participants choose between. There are twelve items for each component.

Short self regulation questionnaire (SSRQ)

The 31 item short self regulation questionnaire was developed by Carey, Neal and Collins (*37*). An example item is "I have trouble making plans to help me reach goals" and responses were on a 5 point scale.

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

The L-Cat (*38*) is a single item that is intended to measure people's 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.

Ten-Item Personality Inventory (TIPI)

Developed by Gosling, Rentfrow and Swann (*39*) 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'), emotional stability (ES; "Calm, emotionally stable'). Participants rate themselves on combinations of two adjectives in each question using a seven point scale.

Theories of Willpower Scale

Developed by Job, Dweck and Walton (*40*) the Theories of Willpower Scale measures people's 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.').

3 factor Eating Questionnaire (TFEQ-R18)

TFEQ-R18 is a shortened measure by Karlsson et al. (*43*) 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.').Questions are presented on four point scales though the options for the scale rating differ across questions.

UPPS-P

Whiteside and Lynam (*33*) initially developed the four factor UPPS after administering a white 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. (*41*) 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 bad consequences’). All items are presented with a four point scale.

Zimbardo Time Perspective Inventory (ZTPI)

ZTPI (*42*) 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’).

Description of Behavioral Tasks

The description of the individual measures borrows text from Enkavi et al. (*9*). Many of the measures have also been described on the Science of Behavior Change's website, and can be demoed there. In addition, the code for individual experiments, which includes information on timing, can be found in the expfactory-experiments repo.
The analysis and post-processing scripts can be found in the expfactory-analysis repo. Data on individual tasks can be found in the Self Regulation Ontology repo.

Adaptive Adjusting Amount Delay Discounting Task

In this task participants make choices between a fixed large amount 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 participant chooses patiently or impatiently in each trial. The amount of adjustments starts at half the immediate amount and is halved at each adjustment. This is repeated for five choices for each fixed later delay and for seven different later delays. The last choice in the procedure is used to estimate the participant’s hyperbolic discount rate (or Effective Delay at 50%). One random trial was chosen and contributed to the total bonus the participant received (note the receipt of this bonus was not linked to their chosen delay in any way).

Behavior was evaluated calculating both a hyperbolic discount rate and area under the (discount) curve for each of the three amounts. We determined the decayed value of the fixed larger amount at each delay using the switch point for the set of seven choices for each delay. These decayed values were both fit a hyperbolic function to calculate the discount rate and to calculate the area under the curve connecting them.

Adaptive N-back Task

In this task participants view a stream of letters on the screen one at a time. They press one button when the letter on the screen matches the letter *N* number of trials ago that is specified at the beginning of each block. They press another button for all other letters. The case of the letters does not matter. Each block consists of twenty plus the load number of letters. The load is increased if the participant has made fewer than three mistakes in the previous block. It is decreased if the participant has made more than five mistakes. Each participant goes through twenty blocks.

To evaluate performance across the whole experiment we calculated the mean load across all blocks. In addition, we used trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with load as a parametric predictor of drift rate.

Angling Risk Task

In this task, which is an extension of the more widely used Balloon Analogue Risk Task (Lejuez et al. 2002), participants play a fishing game for thirty rounds in two conditions. In each round their goal is to catch as many red fish as they can, which translate to earnings in that round. There is also one blue fish in each round; if they catch the blue fish, the round ends and they lose all points for the round. They can end the round whenever they want before catching a blue fish to cash out their earnings for the round.

In the original task, there were two weather conditions: a "sunny" condition where participants could always see how many fish there were in the lake, and a "cloudy" condition where 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 release rules. In the "keep" condition each red fish the participants caught stayed out of the lake (sampling without replacement and increasing the probability of catching a blue fish after each draw). In the "release" condition the red fish were thrown back in the lake so the number of fish in the lake remained constant for the whole round. 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.

We calculated three DVs for each release condition: the adjusted number of clicks (number of clicks on rounds when the blue fish was not caught), the "coefficient of variation" (defined as the standard deviation of the number of clicks on each round when the blue fish was not caught) and the total score in the game. Adjusted clicks and total score were highly correlated, so total score was dropped (see "Data Cleaning and Imputation").

Attentional Network Task

In this task participants 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 before the presentation of the target stimuli: In "no cue" trials no cue is presented before the target stimulus. In "double cue" trials two simultaneous cues are flashed above and below the fixation cross. In "center cue" trials the cue is flashed in the location of the fixation cross. In "spatial cue" trials the cue is flashed in the location where the target stimulus will follow. The cue is a quick flash of a star. Participants complete 24 practice trials and 144 experimental trials (2 (locations) x 4 (cues) x 2 (direction) x 3 (flanker) x 3 (blocks)).

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with cue-type (informative spatial, double, center, and no cue) and flanker-type (congruent, incongruent) as categorical predictors of drift rate. Differences in drift rate coefficients across conditions provides putative measures of three pillars of attention: alerting (no cue - double cue), orienting (central cue - spatial cue) and executive control (incongruent - congruent flanker). In each case, drift rate is expected to be smaller in the former condition, and greater in the latter condition (e.g. "incongruent - congruent drift rate" is generally negative), analogous to a longer reaction time in the former condition.

Choice Reaction Time

In this task participants see either orange or blue squares on the screen for each trial. They are instructed to respond using a different button for each stimulus as quickly and accurately as possible. They complete twenty practice trials and three blocks of fifty test trials.

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift, threshold and non-decision time) using HDDM fit across all participants.

Cognitive Reflection Task

In the classical version of the task participants answer three questions that have numeric answers. The questions are worded such that there is a spontaneous, intuitive but erroneous answer and a correct answer that typically requires a slower and more thoughtful response. Because our sample was likely familiar with the questions of the classical version (Chandler et al. 2014) we have used three items from Toplak, West and Stanovich's (*44*) as well as three from Primi et al.'s (*45*) expansions. Two DVs were calculated from this task: the proportion of correct choices, and the proportion of "intuitive" (but incorrect) choices.

Columbia Card Sorting Task

In this task participants play a card game in multiple rounds. Their goal in each round is to collect as many points as they can by flipping cards from a deck of 32. Each deck contains gain and loss cards. The participants gain points for each gain card they choose, and lose points and immediately end the round if a loss card is chosen. Each gain card is worth either 10 or 30, each loss card costs either 250 or 750 and there are 1 or 3 three loss cards in the round. All the round information is always on display throughout the round. Participants play 24 rounds in two conditions. In the hot condition they flip each card individually and see the outcome of the card immediately whereas in the cold condition they only indicate how many card they would want flip given the round information. Three random trials were chosen which contributed to the overall bonus the participant received.

The number of cards chosen each round was modeled as a function of the amount each gain card was worth, the amount lost if the loss card was chosen, and how many loss cards existed. The standardized beta coefficients for these three variables were taken as sensitivity to gain, loss and probability, respectively. A summary metric of "information use" was also calculated ranging from 0-3, which was the total number of significant ($p < .05$) sensitivity beta coefficients. Finally we also included the average number of cards chosen across all rounds.

Delay Discounting Titrator

In this tasks participants 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. The later reward can be either two or four weeks later than the sooner reward. The sooner amounts are drawn from a normal distribution with a mean of 20 and standard deviation of 10, clipped at 5 and 40. The relative difference between the sooner and later reward can be 1, 5, 10, 15, 20, 25, 30, 50, 75% higher. Participants make 36 choices. One random trial was chosen and contributed to the total bonus the participant received (note the receipt of this bonus was not linked to their chosen delay in any way).

Behavior from this task was evaluated by both tallying the number of patient choices across all trials and fitting a hyperbolic model to the choices where the subjective value of the delayed amount decreases according to the following function: amount/1+discount rate*delay.

Dietary Decision-making Task

This tasks consists of two phases. In the first phase participants 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 chosen. 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).

This preference response was modeled as a function of the current item's health and taste ratings. The standardized coefficients for health and taste were taken as measures of "health sensitivity" and "taste sensitivity" and were the two DVs used for this task.

Digit Span

In this task participants view a series of digits in each trial and are asked to enter them in the order they have seen to a number pad after the digits are presented using the mouse. Participants first complete fourteen trials reporting the digits in the order they have seen and 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. The forward and reverse span were used as the two DVs for this task.

Directed Forgetting Task

In this task participants are presented with six letters forming two rows in each trial. After a 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 participants 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), "negative" (the letter is not in the memory set, but was in the previous trial's memory set) or "control" (the letter is not in the memory set and was not shown in the previous trial). The "negative" trials are intended to create proactive interference between the previous trial and the current trial. Participants completed three rounds of twenty four trials.

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with probe type (positive, negative, control) as a categorical predictor of drift rate. The difference in drift rate coefficients between negative and control probes (negative - control) is putatively related to proactive interference. Drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Dot Pattern Expectancy Task

In this task (MacDonald et al., 2005), which is an adaptation of the AX continuous performance task (Rosvold, Mirsky, Sarason, Bransome, & Beck, 1956; Servan-Schreiber, Cohen, & Steingard, 1996), 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", "BY", or "AY"). 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.

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. 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.

Go/no-go Task

In this task participants see one of two colored squares. They are instructed to respond as quickly as possible 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 stimuli. d` and bias were calculated as the two DVs for this task.

Hierarchical rule learning Task

In this task participants respond to eighteen different stimuli (varying on 3 shapes, 3 orientations and 2 colors) using one of three buttons. In the original work (*46*) there are two rule sets. In the flat rule set each stimulus response pairing has to be learned individually. In the hierarchical rule set a hierarchical relationship between the stimuli and the correct responses allows a two-step policy where e.g. the color indicates whether the response should depend on the shape or the orientation to be a more efficient strategy. We only included the hierarchical rule set in our implementation. There were 360 trials per rule set. Total score was the only DV calculated and contributed to the bonus the participants received at the end of the experiment.

Holt and Laury Titrator

In this task participants choose between two gambles for ten questions. One of the gambles is the safe gamble where the two outcomes have low variance ($80 and $100) and the other gamble is the risky gamble where 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.

We calculated four dependent variables from this task. First we tallied the number of safe choices across the ten gambles. Then we fit the Cumulative Prospect Theory (CPT) as outlined in Toubia et al. (*47*) to extract three parameters: a risk aversion parameter indicating the curvature of the value function, a probability weighting parameter indicating the curvature of the probability weighting function and an inverse temperature parameter indicating how much the behavior uses CPT versus random choice.

Information Sampling Task

In this task participants are presented with a five by five grid of gray boxes. Each box covers one of two colors. Participants were instructed to indicate which color they think is in the majority (one color made up between 13 and 18 of the boxes). To make this decision they can reveal the color of any box by clicking on them. There are two conditions. In the fixed win condition participants 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. Participants complete ten rounds of each round. The DVs from this task are for the average response latency of opening a box (motivation) and the average probability of making the correct decision in each round (see (*48*) for derivation) for each condition.

Keep Track Task

In this task participants are presented with a stream of fifteen words in each round where each word exclusively belongs to one of six categories. 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. The rounds differ in their difficulty based on the number of categories (ranging from 3-5). 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. Each round begins by specifying which categories are relevant that round and participants complete three rounds each for three difficulty levels. 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 total score was the only DV for this task.

Kirby Delay Discounting Items

This 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 participants make choices between smaller immediate monetary amounts and larger delayed monetary amounts. The stimuli are grouped into three (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 participant (00.016-0.025) 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 (note the receipt of this bonus was not linked to their chosen delay in any way).

The performance from this task was evaluated using two metrics. First we tallied the number of patient choices both for all of the trials as well as for each amount group. Then we calculated the hyperbolic discount rate implied by the switch points for each of the three amount group as well.

Local-global Task

In this task 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) letter. 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 letter 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.

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with condition (global vs local), conflict condition (congruent, incongruent, neutral) and switch condition (whether global/local condition was the same or different as the last trial) as categorical predictors of drift rate. Differences in drift rate between global and local conditions putatively relates to a "global bias" (global - local), differences between conflict conditions reflect a general conflict effect (conflict - non-conflict), and differences between stay and switch trials reflect a task-set switch cost (switch - stay). In each case, drift rate is expected to be smaller in the former condition, and greater in the latter condition (e.g. "switch - stay drift rate" is generally negative), analogous to a longer reaction time in the former condition.

Motor Selective Stop Signal Task

In this task 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) appears around the stimulus as the participant prepares their response. Participants are instructed to withhold their response if they see this red star before they respond and if the correct response is either the left or right hand (called the critical hand and randomized across participants). The delay after which the stop signal appeared (stop signal delay) was adjusted using a one-up, one-down staircase procedure in 50ms increments. Participants completed 5 blocks of 60 trials each. 60% of the trials were "go" trials, 20% were "stop" trials (where the stop signal was shown for the critical hand), and 20% were "ignore" trials (where the stop signal was shown for the non-critical hand).

Stop signal reaction time was calculated based on the "critical" trials, a measure that putatively reflects inhibitory control. Using trial-by-trial reaction time and accuracies on the "go" trials, we also calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with "critical condition" (critical vs non-critical hand) and non-stop conditions ("go" vs "ignore") as categorical predictors of drift rate. Proactive control was defined as the difference in drift rate between the two critical conditions (critical - non-critical). Reactive control was defined as the difference in drift rate between the non-stop conditions (ignore - go). In each case, drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Probabilistic Selection Task

This task is divided into two stages. In the first, participants learn to choose between three pairs of abstract shapes based on their reward probabilities. The probabilities for the shapes in each pair are 80%/20%, 70%/30% and 60%/40%. Each learning block is 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%).

Two DVs were 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 right choice) during the test phase using the following formula:

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

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.

Psychological Refractory Period Task

In the psychological refractory period (PRP) task participants respond to two sequential cues (a colored box is displayed, followed by a number). First they respond using one of two buttons depending on the color of a box. Then they 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. Participants completed 32 trials of practice with feedback and 200 test trials without feedback.

The principal effect of interest in the PRP task relates to the idea that processing of the first task slows processing of the second task (either because of a computational "bottleneck" or shared, limited resources; (*49*)). The PRP effect is the observation that the relationship between reaction time on the second task and the ISI approaches a slope of -1 at short ISI's, implying that no additional benefit is gained from additional exposure time to the cue. We calculated the slope between the second task's reaction time and ISI and used that as our only DV. An unsigned (absolute) slope of less than one could be interpreted as reflecting less resource constraint (i.e., enhanced parallel processing).

Raven's Progressive Matrices

Raven's Progressive Matrices (*50*) is a common measure of intelligence, specifically fluid intelligence, that is thought to reflect the ability to infer abstract rules and reason about them to solve problems. In each trial participants are asked to choose the item that would complete a pattern. There were 18 items which increase in difficulty. Total number correct was the only DV.

Recent Probes Task

In this task participants are presented with six letters displayed in two rows. Following the presentation of this memory set participants are presented with a single letter and asked to indicate whether the single letter was in the memory set using one of two buttons. Participants complete twenty four trials per run for three runs. Half of each memory set is from the previous memory set while the other half is novel. The probes were of four types: member of current memory set but not of last two memory sets (positive-not-recent), member of current memory set and of previous memory set (positive-recent), member of previous memory set but not of current memory set (negative-recent) and member of neither of the last two memory sets (negative-not-recent).

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with probe type (positive-recent, positive-not-recent, negative-recent, negative-not-recent) as a categorical predictor of drift rate. Differences in drift rate coefficients between the negative conditioned (negative-recent - negative-not-recent) was taken as a measure of proactive interference. Drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Shape Matching Task

In this tasks participants 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 image 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. Participants 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.

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM 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. Drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Shift Task

In this task 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.

The simplest DV was the overall accuracy on the task (chance being 33%). 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".

The RL model from (*51*) was used to model trial-by-trial performance. This model learns an "attention" weight to different features which is updated based on feedback and informs future choices. These attention weights decay over time. Three DVs were extracted from this model: β (inverse temperature) from the softmax decision function, η (learning rate) for the attention weight updates and *d* (decay rate) for the attention weights.

Simon Task

In this task participants responded using one of two arrow buttons depending on the color of the box they saw on the screen. In the congruent condition the side of the screen matched the response button, in the incongruent condition it did not. Participants completed fifty trials for each condition.

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with "simon condition" (whether the stimulus was on the same side as the response arrow) as a categorical predictor of drift rate. The simon task is primarily analyzed in terms of this effect: response are faster when the stimulus is on the same side as the response key. Differences in drift rate between simon conditions (congruent - incongruent) was the measure of the "simon effect".

Simple Reaction Time

In this task participants are instructed to respond as quickly as possible when they see an "X" on the screen. They complete three blocks of fifty trials. Average reaction time is the only DV.

Spatial Span

In this task participants see a grid of squares in each trial. A sequence of squares is flashed red in each trial. Participants are asked to indicate the sequence that flashed in the order they have seen them 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. The sequence length 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. The forward and reverse span were used as the two DVs for this task.

Stimulus Selective Stop Signal Task

In this task 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 (where the stop signal was shown for the critical hand), and 20% were "ignore" trials (where the stop signal was shown for the non-critical hand).

Stop signal reaction time was calculated based on the "go" and "stop" trials, a measure that putatively reflects inhibitory control. Using trial-by-trial reaction time and accuracies on the "go" trials, we also calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with non-stop conditions ("go" vs "ignore") as a categorical predictor of drift rate. Reactive control was defined as the difference in drift rate between the non-stop conditions (ignore - go). Drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Stop Signal Task

In this task 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). Participants completed 5 blocks of 60 trials each for each condition (the order of the two conditions was randomized across participants).

Stop signal reaction time was calculated separately for each condition. Proactive SSRT speeding was also calculated as the difference in SSRT between the two conditions (20% - 40%). Using trial-by-trial reaction time and accuracies on the "go" trials, we also calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with condition (20% vs 40%) as a categorical predictor of drift rate and threshold. We allows threshold to change as a function of condition because the stop frequency condition was a blocked change (rather than restricted to a particular trial), potentially causing strategic shifts in decision processing, reflected by a changed threshold. Proactive slowing was calculated as the difference in drift rate and threshold between the two conditions (40% - 20%). Drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Stroop

In this task participants were instructed to respond using one of three keys depending on the ink color of the word they were presented. In the congruent condition the word matched the ink color and in the incongruent condition they conflicted. 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).

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with condition (whether the stimulus color was congruent with the word) as a categorical predictor of drift rate. The stroop task is primarily analyzed in terms of this effect: response are faster when the stimulus ink color is the same as the word. Differences in drift rate between congruent and incongruent trials (incongruent - congruent) was the measure of the stroop effect. Drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Cue/Task-switching Task

In this task 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 they make using one of two buttons. 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 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 either switch to the last task (e.g. "color" -> "parity" -> "color") or to a new task (e.g. "color" -> "parity" -> "magnitude"). These three task switch types were balanced across the trials. The "task stay" trials were further subdivided into "cue switch" and "cue stay". 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.

Using trial-by-trial reaction time and accuracies, we calculated individual DDM parameters (drift rate, threshold and non-decision time) using HDDM fit across all participants. In addition, we fit the HDDM with cue condition (switch or stay), task condition (switch or stay) and CTI (100ms or 900ms) as categorical predictors of drift rate. Note that any time there is a task-switch there is a cue-switch. Differences in drift rate based on the cue condition (switch - stay) and task condition (switch - stay) were used as additional DVs. In both cases, drift rate is expected to be smaller in the former condition, and greater in the latter condition, analogous to a longer reaction time in the former condition.

Tower of London

In this task participants are presented with two boards displaying three balls on three pegs: their board and a target board. Participants' task is to make their 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 moves before moving any of them. Each trial is capped at 20 seconds. Participants complete 12 trials of increasing difficulty (the optimal number of moves varied from 2 to 5).

Four DVs were calculated based on this task: average planning time (the time before the first move is initiated), average movement time (the average trial time excluding planning time), number of optimal solutions, and number of extra moves made beyond the optimal number.

Two-step Task

In this task 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.

Three DVs 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 using the lme4 R package (*52*) with the full interactive model fit as a random effect across participants. Individual DVs were defined based on these random effects.

Writing Task

In this task participants 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.

This text was minimally analyzed. We used a sentiment API created at text-processing.com to evaluate the text. This returned a probability of a "positive" and "neutral" classification. Though it is not clear what the exact relationship is between classification probability and intensity, we used these probabilities as our only two DVs extracted from this task.

Tables

Table S1: Self-Report Surveys

Self-Report Surveys	**Dependent Variables**	**References**
BIS-11	-Attentional -Motor -Non-Planning	(*22*)
BIS-BAS	-BAS Drive -BAS Fun-Seeking -BAS Reward-Responsiveness -BIS	(*21*)
Brief Self-Control Scale	-Self-Control	(*23*)
Dickman's Impulsivity Inventory	-Functional	(*24*)
DOSPERT (EB/RP/RT)	-Ethical -Financial -Health/Safety (note: EB[1]) -Recreational -Social	(*25*)
Three-Factor Eating Questionnaire (R18)	-Cognitive Restraint -Emotional Eating -Uncontrolled Eating	(*43*)
Emotion Regulation Questionnaire	-Reappraisal -Suppression	(*26*)
Five Facet Mindfulness Questionnaire	-Acts with Awareness -Describe -Non-Judgment -Non-Reactive -Observe	(*27*)
Future Time Perspective	-Future-Time Perspective	(*28*)
Grit Scale	-Grit	(*53*)
Impulsive-Venturesome Survey	-Impulsiveness[1] -Venturesomeness	(*30*)

Stanford Leisure-Time Activity Categorical Item (L-Cat)	-Activity Level	(*38*)
Mindful Attention Awareness Scale	-Mindfulness	(*31*)
Multidimensional Personality Questionnaire (Control subscale)	-Control[2]	(*32*)
Selection Optimization Compensation	-Elective Selection -Loss-based Selection -Compensation -Optimization[2]	(*36*)
Short Self-Regulation Survey	-Control	(*37*)
Sensation Seeking Survey	-Boredom Susceptibility -Disinhibition -Experience Seeking -Thrill/Adventure Seeking	(*34*)
Ten Item Personality Questionnaire	-Agreeableness -Conscientiousness -Emotional Stability -Extraversion -Openness	(*39*)
Theories of Willpower	-Endorse Limited Resource	(*40*)
Time Perspective Survey	-Future -Past Negative -Past Positive -Present Fatalistic -Present Hedonistic	(*42*)
UPPS+P	-Lack of Perseverance -Lack of Premeditation -Negative Urgency -Positive Urgency -Sensation Seeking	(*41*)

[1] Log transformed due to high positive skew (skew > 1)

[2] Reflected and log transformed due to high negative skew (skew < -1)

Table S2: Behavioral Tasks

Task	**Dependent Variables**	**References**
Adaptive N-Back	DDM Parameters[1] Drift Rate as a function of load Average load	(*54*, *55*)
Angling Risk Task	Two Conditions (Keep, Release): Adjusted Clicks Coefficient of Variation Score[2]	(*56*, *57*)
Attention Network Task	DDM Parameters[1] Alerting Effect Orienting Effect Conflict Effect	(*58*)
Bickel Titrator	Discount Rate for two payout magnitudes[3]	(*59*)
Choice Reaction Time	DDM Parameters[1]	
Cognitive Reflection Task	Correct Proportion Intuitive Proportion	(*44*, *45*)
Columbia Card Task Cold/Hot	Average # of cards chosen Gain Sensitivity Loss Sensitivity # Loss Cards Sensitivity Level of Information Use	(*60*)
Dietary Decision Task	Health Sensitivity Taste Sensitivity	(*61*)
Digit Span	Forward Span Reverse Span	(*62*)
Directed Forgetting	-DDM Parameters[1] -Proactive Interference	(*63*)
Discount Titrator	-Percent Patient	(*64*)
Dot Pattern Expectancy	-DDM Parameters[1] -AY-BY -BX-BY -D-prime -Bias	(*65*)

Go-NoGo	-D-prime -Bias	
Hierarchical Learning Task	-Total Score	(*46*)
Holt & Laury	-Percent Patient -Beta (inverse softmax temperature) -Risk Aversion (value function curvature) -# Safe Choices	(*66*)
Information Sampling Task	Two conditions (Decreasing Win, Fixed Win): -Probability Correct at choice -Motivation	(*48*)
Keep Track Task	-Score	(*67*, *68*)
Kirby	-Discount Rate for three payout magnitudes[3] -Percent Patient Choices -Percent Patient Choices for three payout magnitudes[2]	(*69*)
Local-Global	-DDM Parameters[1] -Switch Cost -Conflict Effect -Global Bias	(*67*, *68*)
Motor Selective Stop Signal	-DDM Parameters[1] -SSRT -Reactive Control -Selective Proactive Control -Proactive Control	(*70*)
Probabilistic Selection Task	-Positive Learning Bias -Value Sensitivity[2]	(*71*)
Psychological Refractory Period	-Slope of PRP function	(*49*)
Raven's Progressive Matrices	-Score	(*50*)
Recent Probes	-DDM Parameters[1] -Proactive Interference	(*63*)
Shape Matching Task	-DDM Parameters[1]	(*72*)

	-Stimulus Interference	
Shift Task	-Accuracy -Learning Rate -Learning to Learn -Model Parameters: - Beta (inverse softmax temperature) - Attentional Decay - RL Learning Rate	(*51*, *73*)
Simon Task	-DDM Parameters[1] -Simon Effect	(*74*)
Simple Reaction Time	-Average Reaction Time	
Spatial Span	-Forward Span -Reverse Span	(*62*)
Stimulus Selective Stop Signal	-DDM Parameters[1] (note: thresh[3]) -SSRT -Reactive Control	(*75*)
Stop Signal	-DDM Parameters[1] -SSRT (low stop signal probability condition) -SSRT (high stop signal probability condition) -Proactive SSRT speeding -Proactive Slowing	(*76*)
Stroop	-DDM Parameters[1] -Stroop Effect	(*67*, *68*)
Cue/Task-Switch	-DDM Parameters[1] -Stimulus Switch Cost -Task Switch Cost	(*77*)
Tower of London	-Average Move Time -# Extra Moves -# Optimal Solutions -Planning Time	(*78*)
Two-step Decision	-Model-Based Index -Model-Free Index -Perseverance	(*79*)

Writing Task	-Sentiment Analysis: -Positive Probability -Negative Probability	

[1] DDM Parameters include drift rate, threshold and non-decision time
[2] Dropped due to high ($r > 0.85$) correlations with another DV in the same measure
[3] Log transformed due to high positive skew (skew > 1)
[4] Reflected and log transformed due to high negative skew (skew < -1)

Table S3: Prediction results using factor scores

	Binge Drinking	Problem Drinking	Unsafe Drinking	Drug Use	Lifetime Smoking	Daily Smoking	Mental Health	Obesity	Income/ Life-outcomes
Task: Ridge[2]	**R^2 = 0.0 (.01)[1] MAE = .79 (.78)**	**R^2 = 0.0 (.01) MAE = .58 (.57)**	**R^2 = 0.0 (.01) MAE = .54 (.53)**	**R^2 = .01 (0.0) MAE = .51 (.50)**	**R^2 = .01 (.02) MAE = .95 (.94)**	**R^2 = .02 (.04) MAE = .82 (.81)**	**R^2 = 0.0 (.01) MAE = .79 (.78)**	**R^2 = .03 (.04) MAE = .85 (.85)**	**R^2 = .04 (.05) MAE = .77 (.76)**
Task: Lasso	R^2 = .01 (.01) MAE = .79 (.79)	R^2 = .01 (0.0) MAE = .58 (.58)	R^2 = 0.0 (0.0) MAE = .53 (.53)	R^2 = .01 (nan) MAE = .50 (nan)	R^2 = .01 (.02) MAE = .95 (.94)	R^2 = .02 (.03) MAE = .83 (.83)	R^2 = 0.0 (.01) MAE = .79 (.78)	R^2 = .02 (.04) MAE = .86 (.86)	R^2 = .03 (.05) MAE = .77 (.76)
Task: Random Forest	R^2 = 0.0 (1.0) MAE = .85 (0.0)	R^2 = 0.0 (1.0) MAE = .66 (0.0)	R^2 = 0.0 (1.0) MAE = .71 (0.0)	R^2 = 0.0 (1.0) MAE = .62 (0.0)	R^2 = 0.0 (1.0) MAE = .95 (0.0)	R^2 = .01 (1.0) MAE = .85 (0.0)	R^2 = 0.0 (1.0) MAE = .84 (0.0)	R^2 = 0.0 (1.0) MAE = .90 (0.0)	R^2 = .01 (1.0) MAE = .81 (0.0)
Task: SVM	R^2 = 0.0 (.01) MAE = .73 (.72)	R^2 = 0.0 (0.0) MAE = .43 (.43)	R^2 = .01 (.01) MAE = .52 (.52)	R^2 = 0.0 (0.0) MAE = .41 (.40)	R^2 = .01 (.01) MAE = .91 (.89)	R^2 = .01 (.03) MAE = .71 (.71)	R^2 = 0.0 (.01) MAE = .75 (.74)	R^2 = .03 (.04) MAE = .73 (.73)	R^2 = .03 (.05) MAE = .76 (.75)
Survey: Ridge[2]	**R^2 = .13 (.17) MAE = .70 (.68)**	**R^2 = .07 (.11) MAE = .57 (.56)**	**R^2 = .10 (.14) MAE = .59 (.58)**	**R^2 = .04 (.08) MAE = .53 (.51)**	**R^2 = .04 (.08) MAE = .91 (.89)**	**R^2 = .04 (.08) MAE = .81 (.78)**	**R^2 = .29 (.32) MAE = .59 (.58)**	**R^2 = .14 (.18) MAE = .77 (.75)**	**R^2 = .04 (.08) MAE = .76 (.74)**
Survey: Lasso	R^2 = .13 (.17) MAE = .70 (.68)	R^2 = .06 (.11) MAE = .56 (.55)	R^2 = .09 (.14) MAE = .59 (.57)	R^2 = .04 (.08) MAE = .51 (.50)	R^2 = .04 (.06) MAE = .94 (.93)	R^2 = .04 (.08) MAE = .81 (.79)	R^2 = .26 (.32) MAE = .61 (.58)	R^2 = .14 (.18) MAE = .78 (.75)	R^2 = .03 (.07) MAE = .77 (.75)
Survey: Random Forest	R^2 = .1 (1.0) MAE = .73 (0.0)	R^2 = .02 (1.0) MAE = .61 (0.0)	R^2 = .02 (1.0) MAE = .69 (0.0)	R^2 = .01 (1.0) MAE = .59 (0.0)	R^2 = .02 (1.0) MAE = .91 (0.0)	R^2 = .02 (1.0) MAE = .85 (0.0)	R^2 = .22 (1.0) MAE = .64 (0.0)	R^2 = .08 (1.0) MAE = .81 (0.0)	R^2 = .01 (1.0) MAE = .83 (0.0)

Survey: SVM	R^2 = .13 (.16) MAE = .68 (.66)	R^2 = .01 (0.0) MAE = .43 (.43)	R^2 = .09 (.12) MAE = .51 (.51)	R^2 = 0.0 (0.0) MAE = .4 (.4)	R^2 = .05 (.07) MAE = .88 (.84)	R^2 = .02 (.07) MAE = .72 (.70)	R^2 = .28 (.31) MAE = .58 (.56)	R^2 = .13 (.17) MAE = .72 (.70)	R^2 = .04 (.07) MAE = .76 (.73)
Task and Survey: Ridge	R^2 = .12 (.18) MAE = .70 (.67)	R^2 = .07 (.13) MAE = .58 (.56)	R^2 = .09 (.15) MAE = .61 (.58)	R^2 = .03 (.08) MAE = .53 (.51)	R^2 = .05 (.09) MAE = .90 (.87)	R^2 = .06 (.11) MAE = .79 (.77)	R^2 = .29 (.34) MAE = .59 (.57)	R^2 = .15 (.21) MAE = .77 (.74)	R^2 = .07 (.13) MAE = .75 (.72)
Task and Survey: Lasso	R^2 = .13 (.18) MAE = .70 (.68)	R^2 = .07 (.13) MAE = .55 (.54)	R^2 = .09 (.14) MAE = .57 (.55)	R^2 = .03 (.08) MAE = .51 (.50)	R^2 = .03 (.07) MAE = .94 (.92)	R^2 = .04 (.09) MAE = .81 (.79)	R^2 = .28 (.33) MAE = .60 (.58)	R^2 = .15 (.20) MAE = .77 (.75)	R^2 = .08 (.13) MAE = .74 (.72)
Task and Survey: Random Forest	R^2 = .06 (1.0) MAE = .75 (0.0)	R^2 = .02 (1.0) MAE = .63 (0.0)	R^2 = .03 (1.0) MAE = .66 (0.0)	R^2 = .04 (1.0) MAE = .55 (0.0)	R^2 = .02 (1.0) MAE = .91 (0.0)	R^2 = .02 (1.0) MAE = .83 (0.0)	R^2 = .23 (1.0) MAE = .63 (0.0)	R^2 = .10 (1.0) MAE = .78 (0.0)	R^2 = .02 (1.0) MAE = .82 (0.0)
Task and Survey: SVM	R^2 = .12 (.17) MAE = .68 (.65)	R^2 = .01 (.02) MAE = .43 (.43)	R^2 = .08 (.12) MAE = .52 (.50)	R^2 = .00 (.01) MAE = .40 (.39)	R^2 = .03 (.08) MAE = .89 (.82)	R^2 = .05 (.09) MAE = .72 (.70)	R^2 = .28 (.31) MAE = .58 (.55)	R^2 = .14 (.19) MAE = .72 (.69)	R^2 = .08 (.12) MAE = .74 (.70)

[1] Insample score is displayed in parentheses.

[2] Bolded values are used in Figures 5,6 in the main text.

Table S4: Prediction results using DVs

	Binge Drinking	Problem Drinking	Unsafe Drinking	Drug Use	Lifetime Smoking	Daily Smoking	Mental Health	Obesity	Income/ Life-outcomes
Task: Ridge	R2 = .00 (.26)[1] MAE = .90 (.67)	R2 = .00 (.24) MAE = .75 (.55)	R2 = .00 (.23) MAE = .78 (.58)	R2 = .00 (.26) MAE = .72 (.53)	R2 = .00 (.27) MAE = .97 (.74)	R2 = .03 (.34) MAE = .89 (.66)	R2 = .00 (.27) MAE = .89 (.66)	R2 = .02 (.30) MAE = .91 (.70)	R2 = .04 (.34) MAE = .84 (.64)
Task: Lasso	R2 = .02 (.03) MAE = .79 (.79)	R2 = .01 (.02) MAE = .58 (.58)	R2 = .01 (.02) MAE = .53 (.53)	R2 = .01 (.02) MAE = .50 (.50)	R2 = .02 (.06) MAE = .97 (.95)	R2 = .01 (.15) MAE = .82 (.78)	R2 = .02 (.04) MAE = .78 (.78)	R2 = .01 (.11) MAE = .87 (.84)	R2 = .11 (.18) MAE = .74 (.72)
Survey: Ridge	R2 = .18 (.35) MAE = .67 (.59)	R2 = .03 (.21) MAE = .64 (.55)	R2 = .08 (.27) MAE = .68 (.59)	R2 = .04 (.20) MAE = .61 (.53)	R2 = .04 (.22) MAE = .89 (.78)	R2 = .06 (.23) MAE = .81 (.71)	R2 = .24 (.42) MAE = .62 (.53)	R2 = .13 (.33) MAE = .78 (.67)	R2 = .06 (.24) MAE = .78 (.68)
Survey: Lasso	R2 = .25 (.27) MAE = .63 (.62)	R2 = .09 (.13) MAE = .54 (.53)	R2 = .10 (.17) MAE = .56 (.54)	R2 = .07 (.10) MAE = .50 (.49)	R2 = .06 (.09) MAE = .92 (.91)	R2 = .06 (.17) MAE = .80 (.76)	R2 = .28 (.35) MAE = .59 (.57)	R2 = .17 (.20) MAE = .77 (.76)	R2 = .08 (.17) MAE = .75 (.71)

[1] Insample score is displayed in parentheses.

Figures

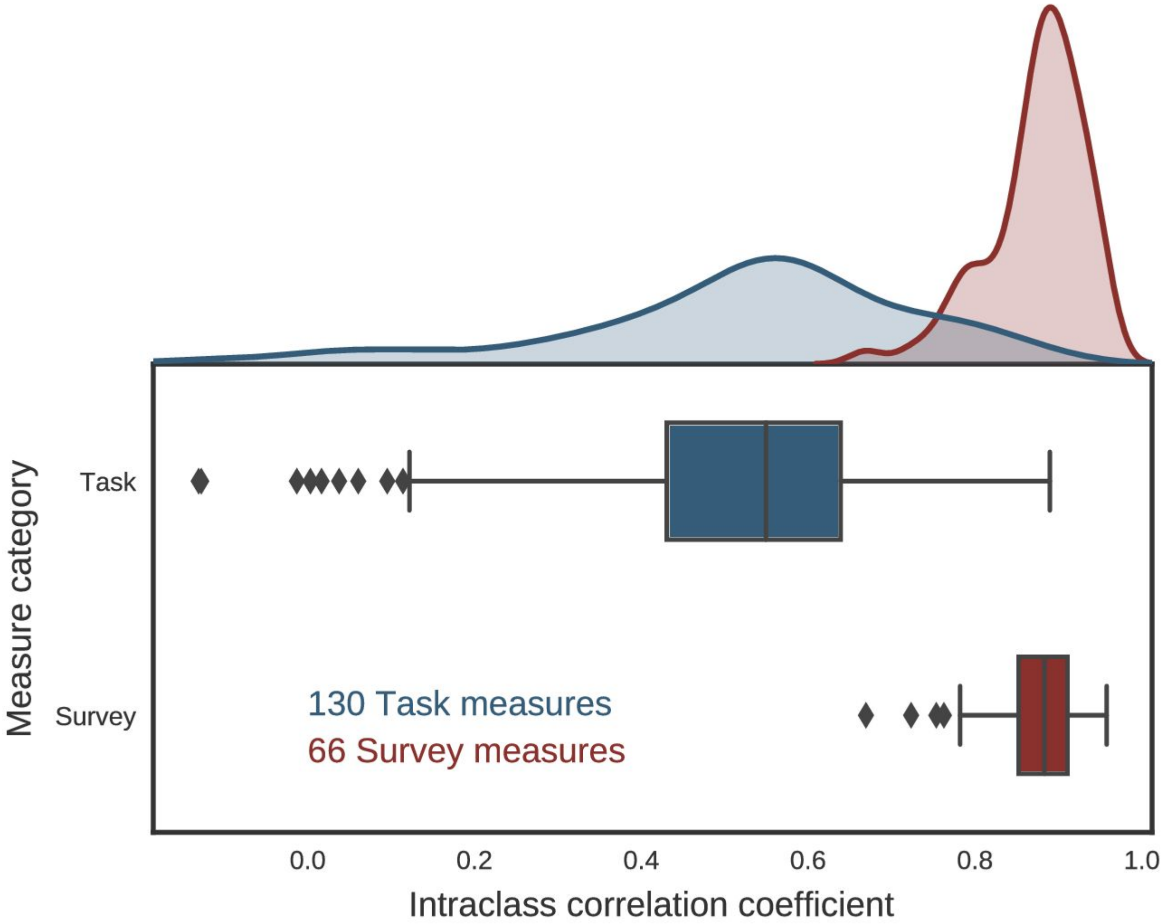


Fig. S1. Test-retest reliability. Test-retest reliability for the survey and task DVs, as quantified by bootstrapped intraclass correlation coefficient. The full procedure is outlined in Enkavi et al. (*9*).

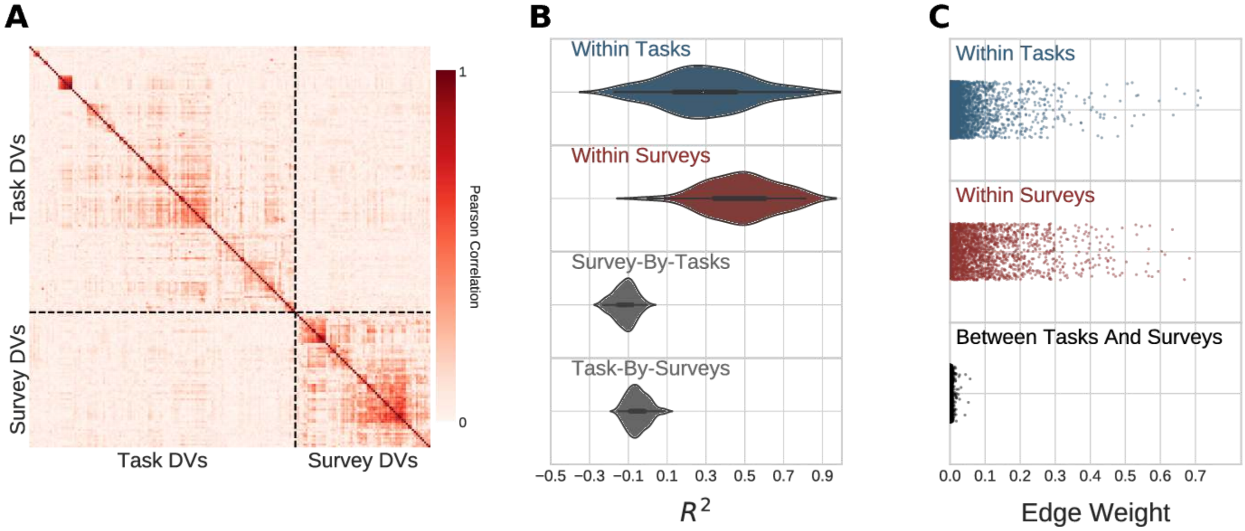


Fig. S2. Survey-task DV relationships. (a) Pearson correlation between task and survey DVs. DVs are organized by measurement category and ordered based on the respective hierarchical clustering solutions. (b) Cross-validated R^2 derived from cross-validated ridge regression of either a single task or survey DV using all survey or task DVs (holding out the target). (c) Estimate of relationships between or across measurement-category according to the graphical lasso used to estimate Figure 2.

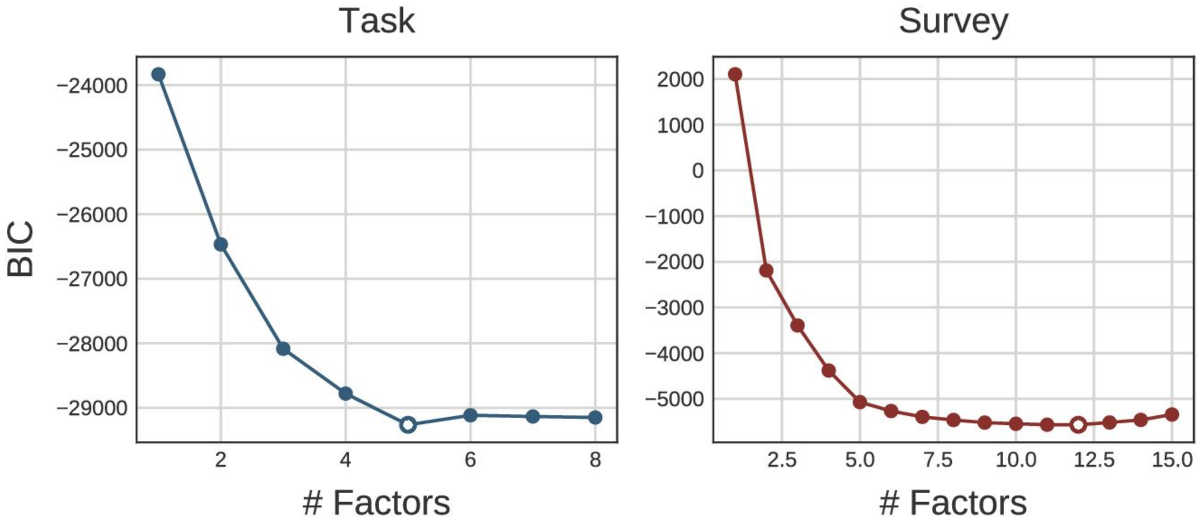


Fig. S3. Bayesian information criterion (BIC) curves for EFA. BIC was used to determine the optimal number of factors to extract for exploratory factor analysis. The BIC values for a range of factors are shown for surveys and tasks. The optimal dimensionality is indicated by an empty circle.

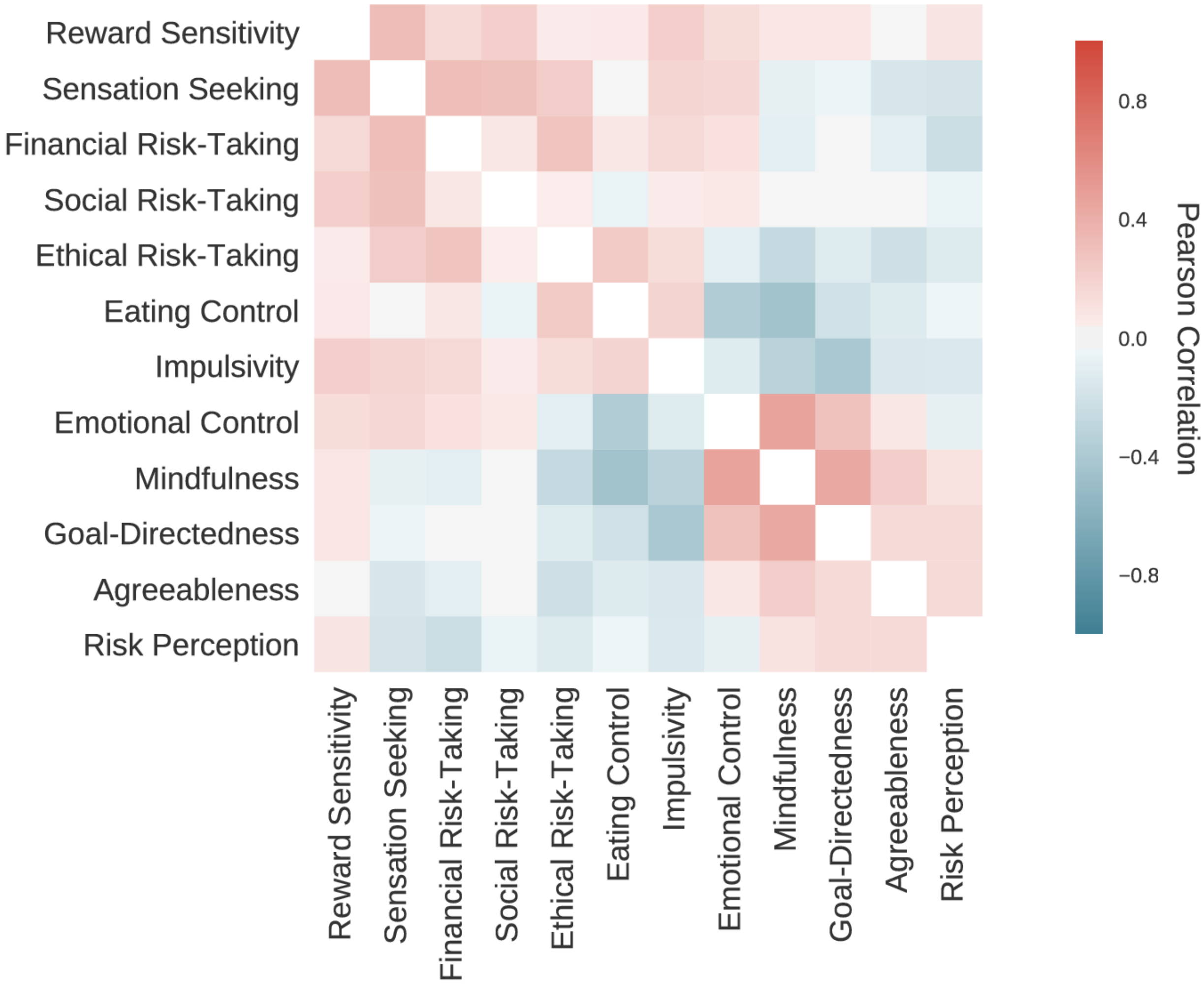


Fig. S4. Survey factor correlations. Exploratory factor analysis was performed with oblimin rotation, which allows correlations between factors. The heatmap of these correlations is shown, with the factors ordered by a hierarchical clustering analysis. Note that the colorbar has been scaled based on correlations between factors - *r* values on the diagonal are all 1.

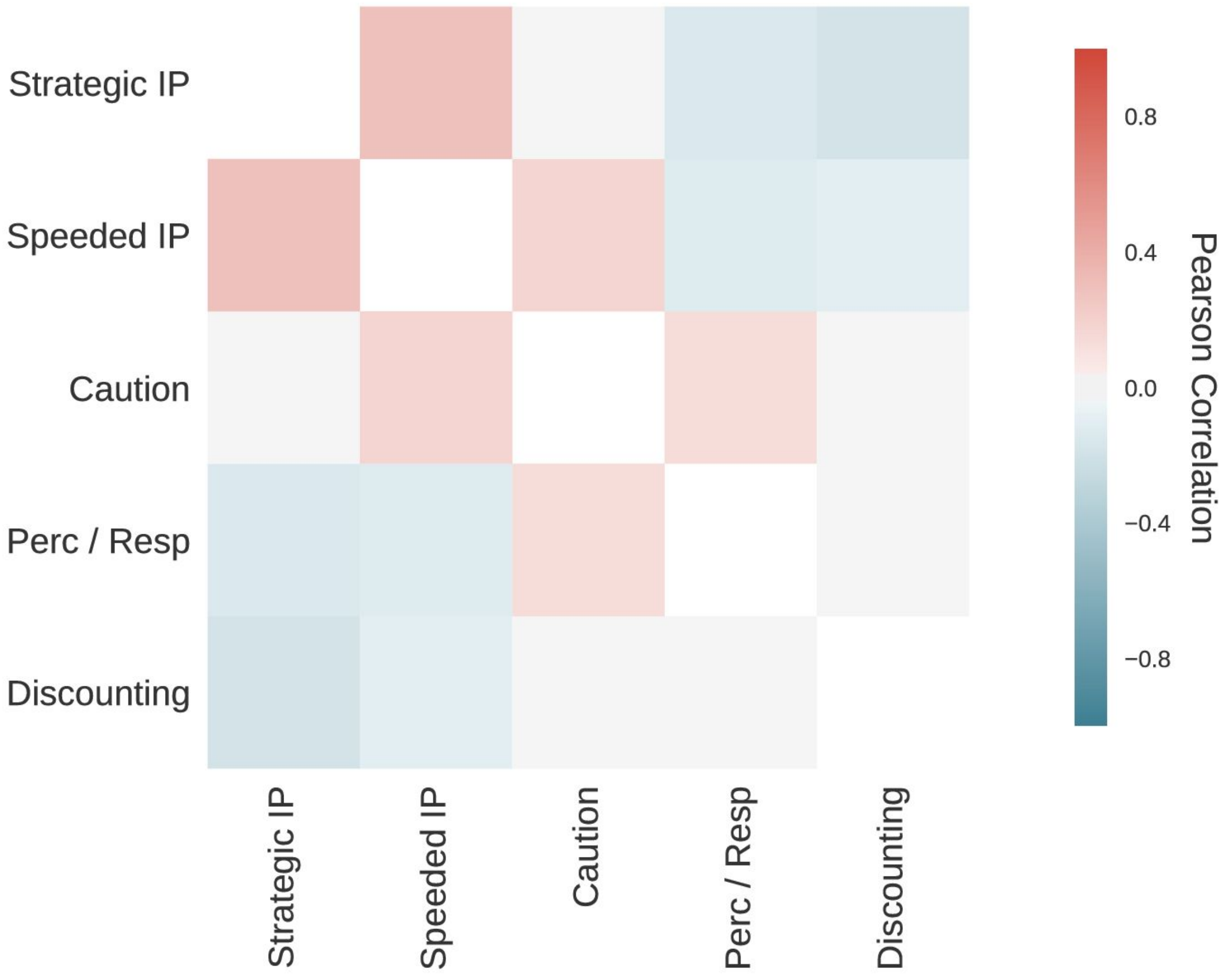


Fig. S5. Task factor correlations. Exploratory factor analysis was performed with oblimin rotation, which allows correlations between factors.The heatmap of these correlations is shown, with the factors ordered by a hierarchical clustering analysis. Note that the colorbar has been scaled based on correlations between factors - *r* values on the diagonal are all 1.

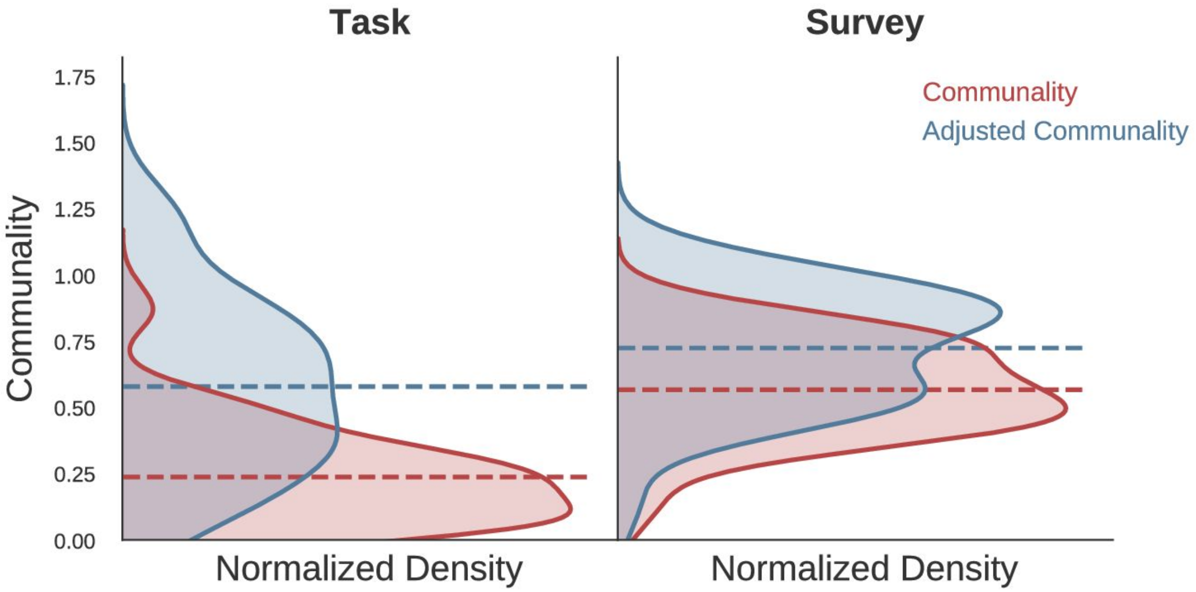


Fig. S6. Communality correction for test-retest reliability. The distribution of communality (the variance explained by the related EFA model) across DVs is shown in red. The average communality (equivalent to the variance explained of the entire measurement category) is depicted with a red dashed line. Communality was adjusted by dividing by the test-retest reliability, as assessed by Pearson correlation, resulting in the blue distribution. Only DVs with a test-retest reliability above .2 are included in these distributions. Note that the surveys EFA model performs better than the task EFA model (red curves, survey R^2 = .58, task R^2 = .26), but this difference is attenuated after adjusting for test-retest reliability (blue curves, survey adjusted R^2 = .72, task adjusted R^2 = .57**)**

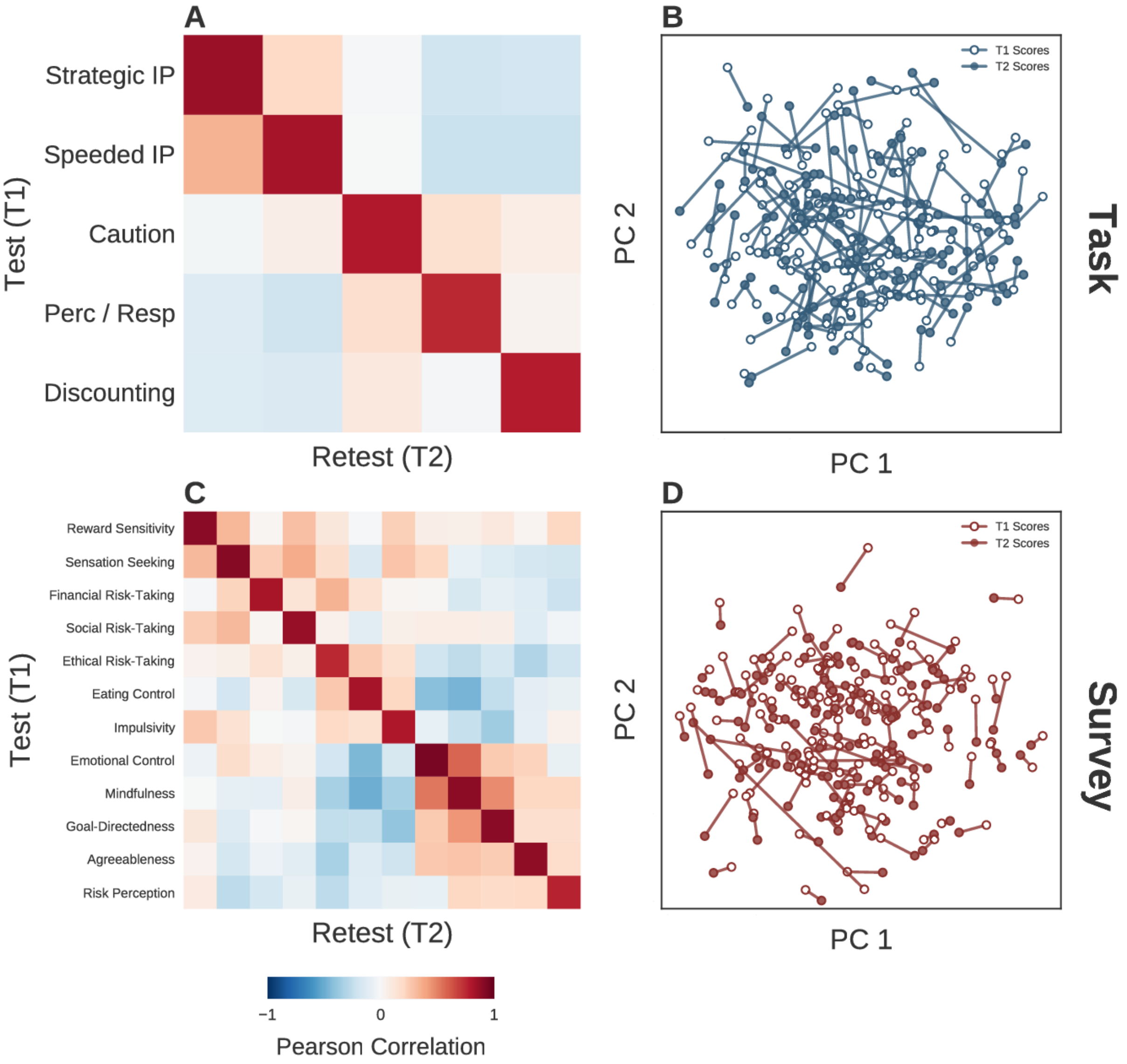


Fig. S7. EFA test-retest reliability. The EFA model derived from the full dataset (n=522) was applied to a subsample of participants who repeated the entire battery within 4 months (n=150) to compute factor scores (T2 scores). These factor scores were then correlated with factors scores derived from the same participants during their original testing (T1 scores). Heatmaps reflecting these correlations are displayed for both tasks (A) and surveys (C). The average Pearson's *r* value between T1 and T2 factor scores was .86 for surveys and .81 for tasks. Also note the similarity of the off-diagonals to Fig S6-7. To visualize the relative stability of individual factor scores compared to group variability the factor scores were projected into a 2-dimensional space defined using PCA on T1 factor scores (B, D). T1 scores are depicted using an empty circle, while T2 scores are depicted using a filled circle, with each individual corresponding to one "stick". It is evident that while factor scores are not perfectly stable, there is substantially more variability across individuals than within - an essential feature of a reasonable individual difference measure.

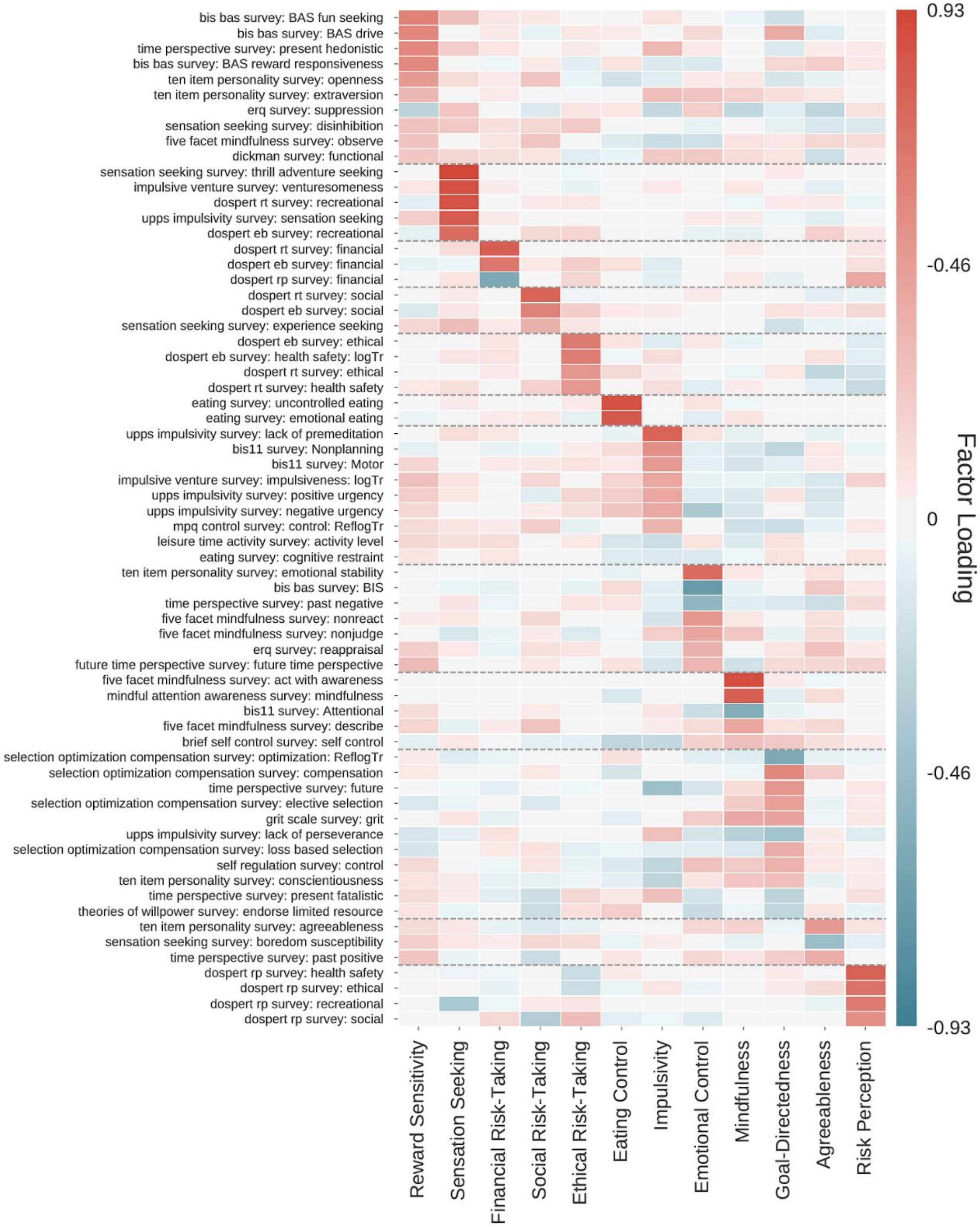
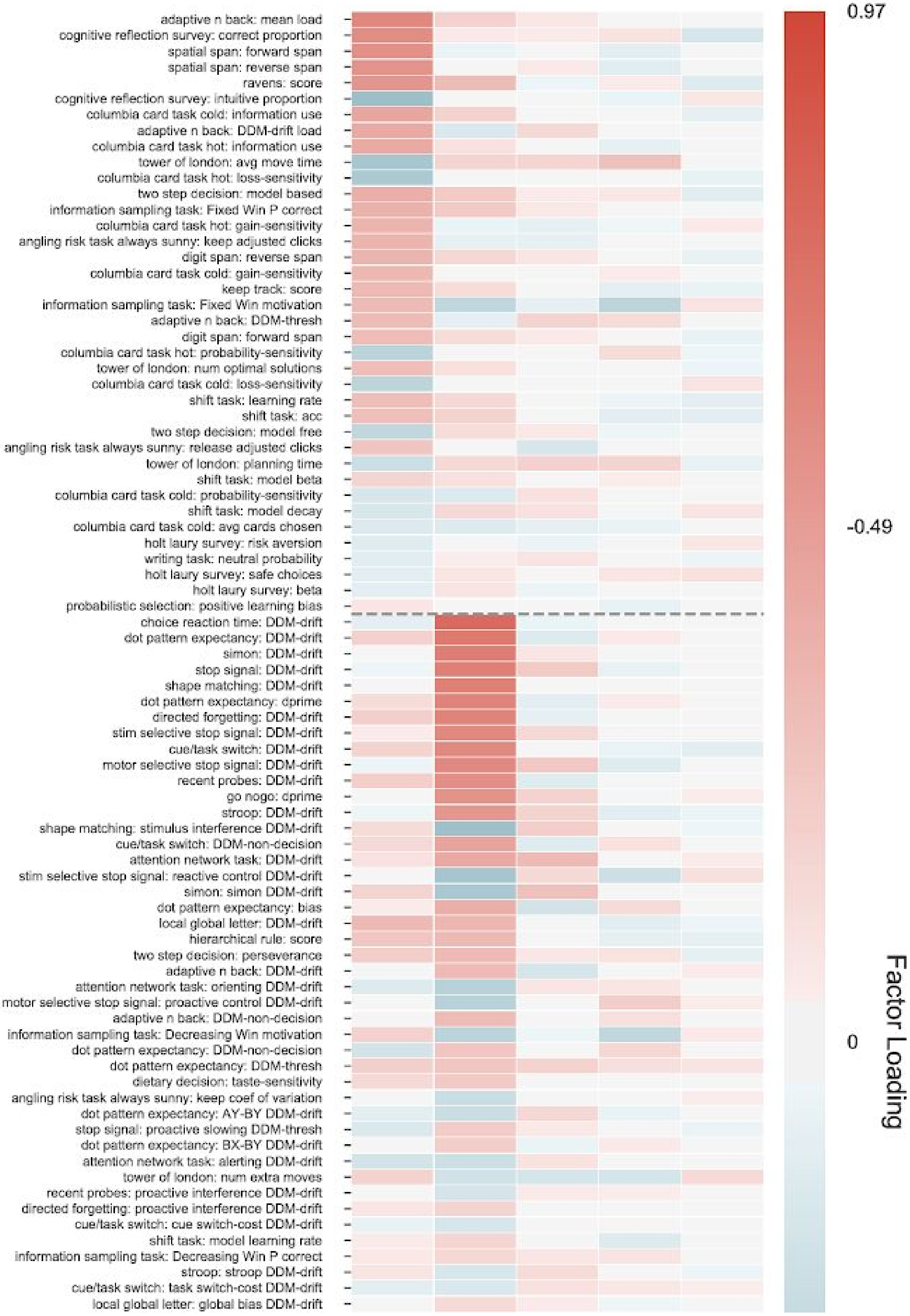


Fig. S8. Survey factor loadings. 12 factors were determined using a BIC criteria for exploratory factor analysis. The 66 survey DVs are grouped and ordered based on the largest (absolute) factor loading for that DV. Dotted lines indicate separate groups derived from this criteria, and are used for visualization purposes only.


adaptive n back: mean load
cognitive reflection survey: correct proportion
spatial span: forward span
spatial span: reverse span
ravens: score
cognitive reflection survey: intuitive proportion
columbia card task cold: information use
adaptive n back: DDM-drift load
columbia card task hot: information use
tower of london: avg move time
columbia card task hot: loss-sensitivity
two step decision: model based
information sampling task: Fixed Win P correct
columbia card task hot: gain-sensitivity
angling risk task always sunny: keep adjusted clicks
digit span: reverse span
columbia card task cold: gain-sensitivity
keep track: score
information sampling task: Fixed Win motivation
adaptive n back: DDM-thresh
digit span: forward span
columbia card task hot: probability-sensitivity
tower of london: num optimal solutions
columbia card task cold: loss-sensitivity
shift task: learning rate
shift task: acc
two step decision: model free
angling risk task always sunny: release adjusted clicks
tower of london: planning time
shift task: model beta
columbia card task cold: probability-sensitivity
shift task: model decay
columbia card task cold: avg cards chosen
holt laury survey: risk aversion
writing task: neutral probability
holt laury survey: safe choices
holt laury survey: beta
probabilistic selection: positive learning bias
choice reaction time: DDM-drift
dot pattern expectancy: DDM-drift
simon: DDM-drift
stop signal: DDM-drift
shape matching: DDM-drift
dot pattern expectancy: dprime
directed forgetting: DDM-drift
stim selective stop signal: DDM-drift
cue/task switch: DDM-drift
motor selective stop signal: DDM-drift
recent probes: DDM-drift
go nogo: dprime
stroop: DDM-drift
shape matching: stimulus interference DDM-drift
cue/task switch: DDM-non-decision
attention network task: DDM-drift
stim selective stop signal: reactive control DDM-drift
simon: simon DDM-drift
dot pattern expectancy: bias
local global letter: DDM-drift
hierarchical rule: score
two step decision: perseverance
adaptive n back: DDM-drift
attention network task: orienting DDM-drift
motor selective stop signal: proactive control DDM-drift
adaptive n back: DDM-non-decision
information sampling task: Decreasing Win motivation
dot pattern expectancy: DDM-non-decision
dot pattern expectancy: DDM-thresh
dietary decision: taste-sensitivity
angling risk task always sunny: keep coef of variation
dot pattern expectancy: AY-BY DDM-drift
stop signal: proactive slowing DDM-thresh
dot pattern expectancy: BX-BY DDM-drift
attention network task: alerting DDM-drift
tower of london: num extra moves
recent probes: proactive interference DDM-drift
directed forgetting: proactive interference DDM-drift
cue/task switch: cue switch-cost DDM-drift
shift task: model learning rate
information sampling task: Decreasing Win P correct
stroop: stroop DDM-drift
cue/task switch: task switch-cost DDM-drift
local global letter: global bias DDM-drift
0.97
-0.49
0
Factor Loading

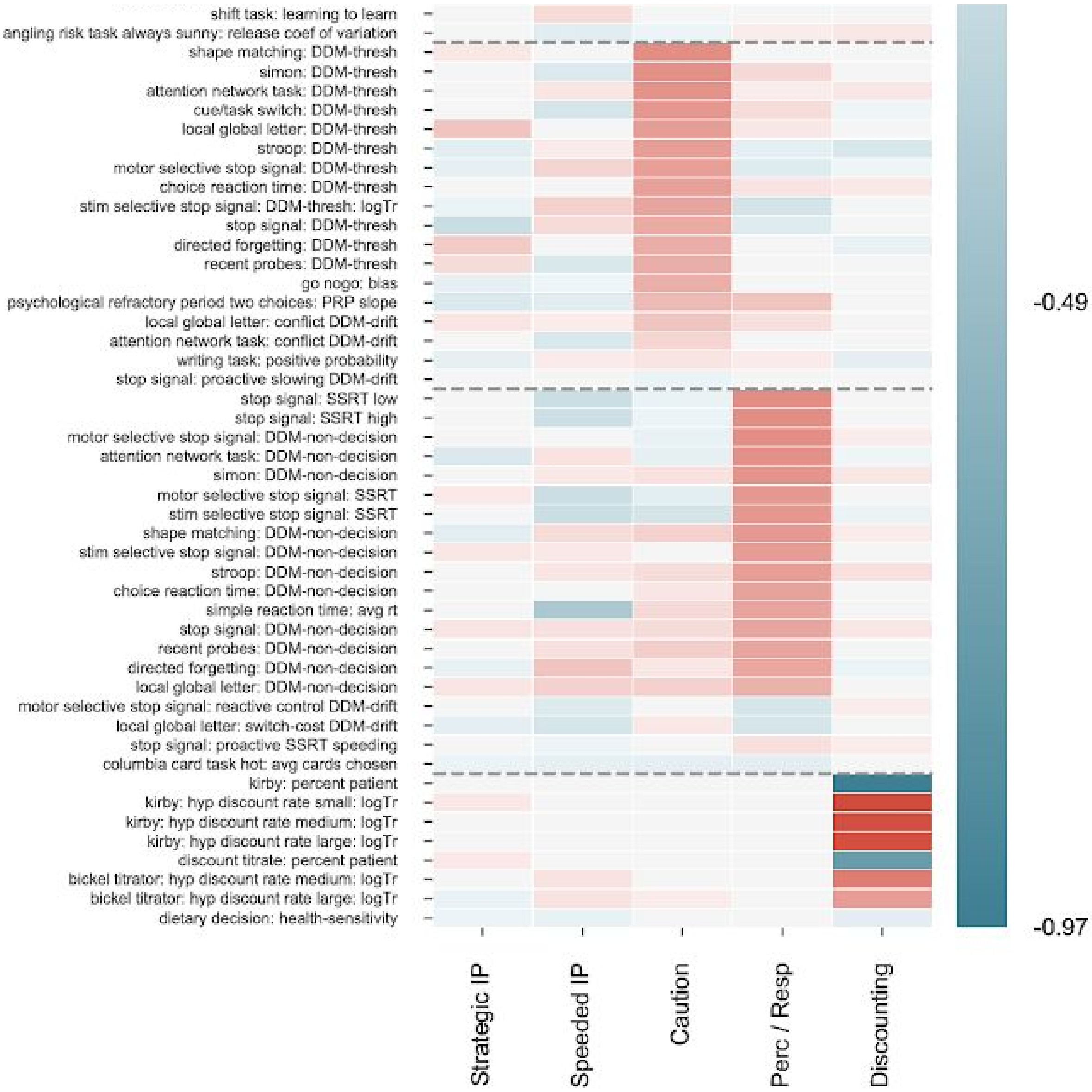


Fig. S9. Task factor loadings. 5 factors were determined using a BIC criteria for exploratory factor analysis. The 130 survey DVs are grouped and ordered based on the largest (absolute) factor loading for that DV. Dotted lines indicate separate groups derived from this criteria, and are used for visualization purposes only.

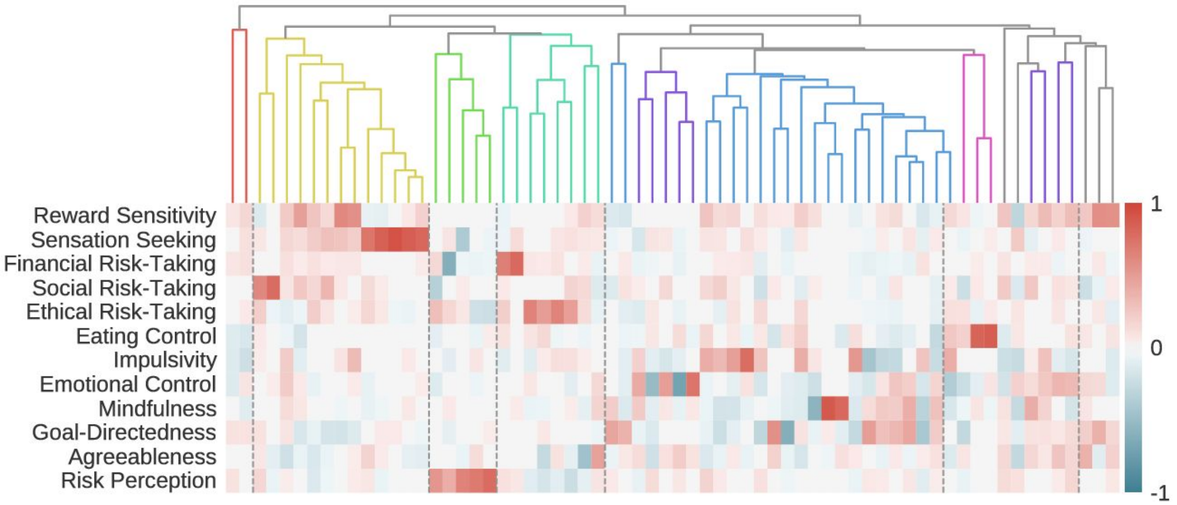


Fig. S10. Survey ontology with clusters in participant space. This figure is identical to Figure 2 in the main paper, except the hierarchical clustering algorithm is operating over participant scores for each DV rather than factor loading. Thus each DV is represented by a 522-dimensional vector. The increased height and lack of clear clustering in the dendrogram shows that DVs are not readily categorized in this "participant" space.

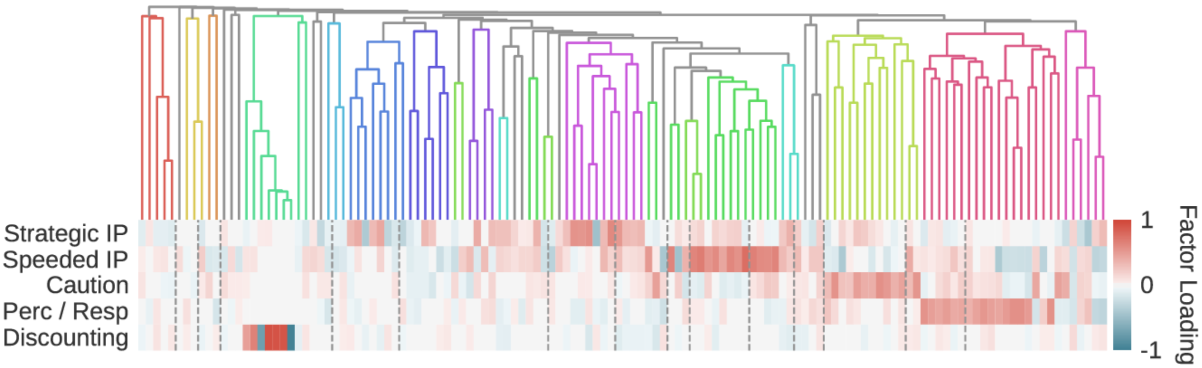


Fig. S11. Task ontology with clusters in participant space. This figure is identical to figure 3 in the main paper, except the hierarchical clustering algorithm is operating over participant scores for each DV rather than factor loading. Thus each DV is represented by a 522-dimensional vector. The increased height and lack of clear clustering in the dendrogram shows that DVs are not readily categorized in this "participant" space.

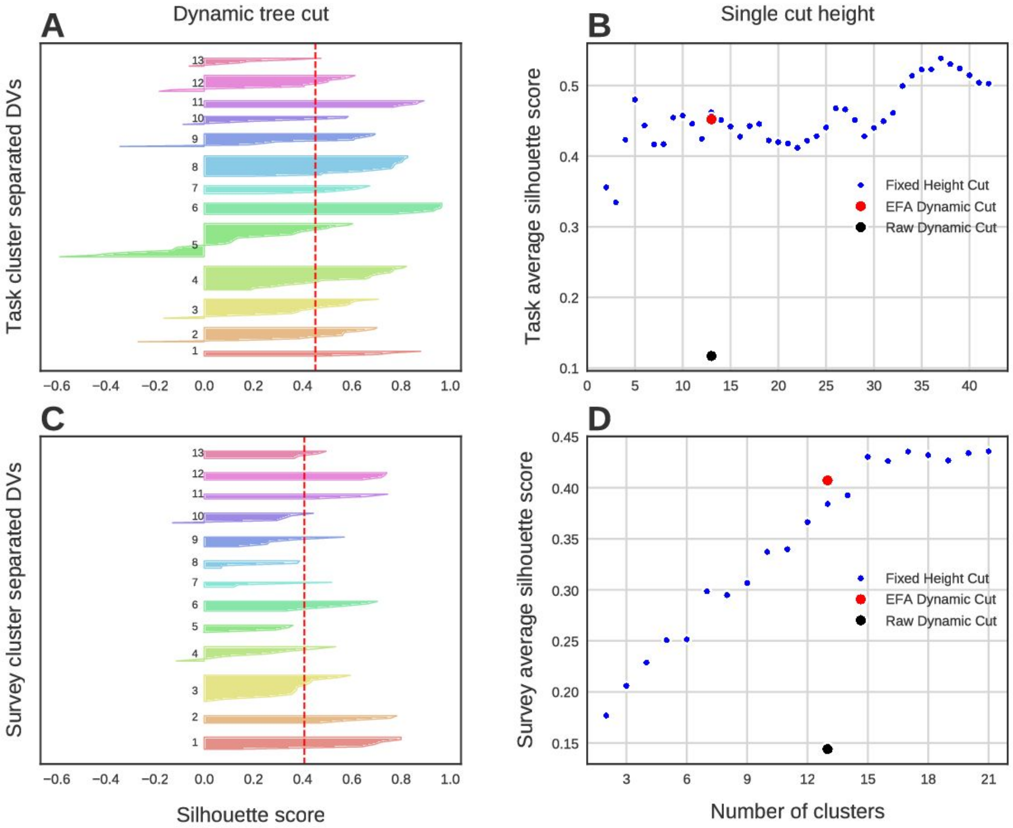
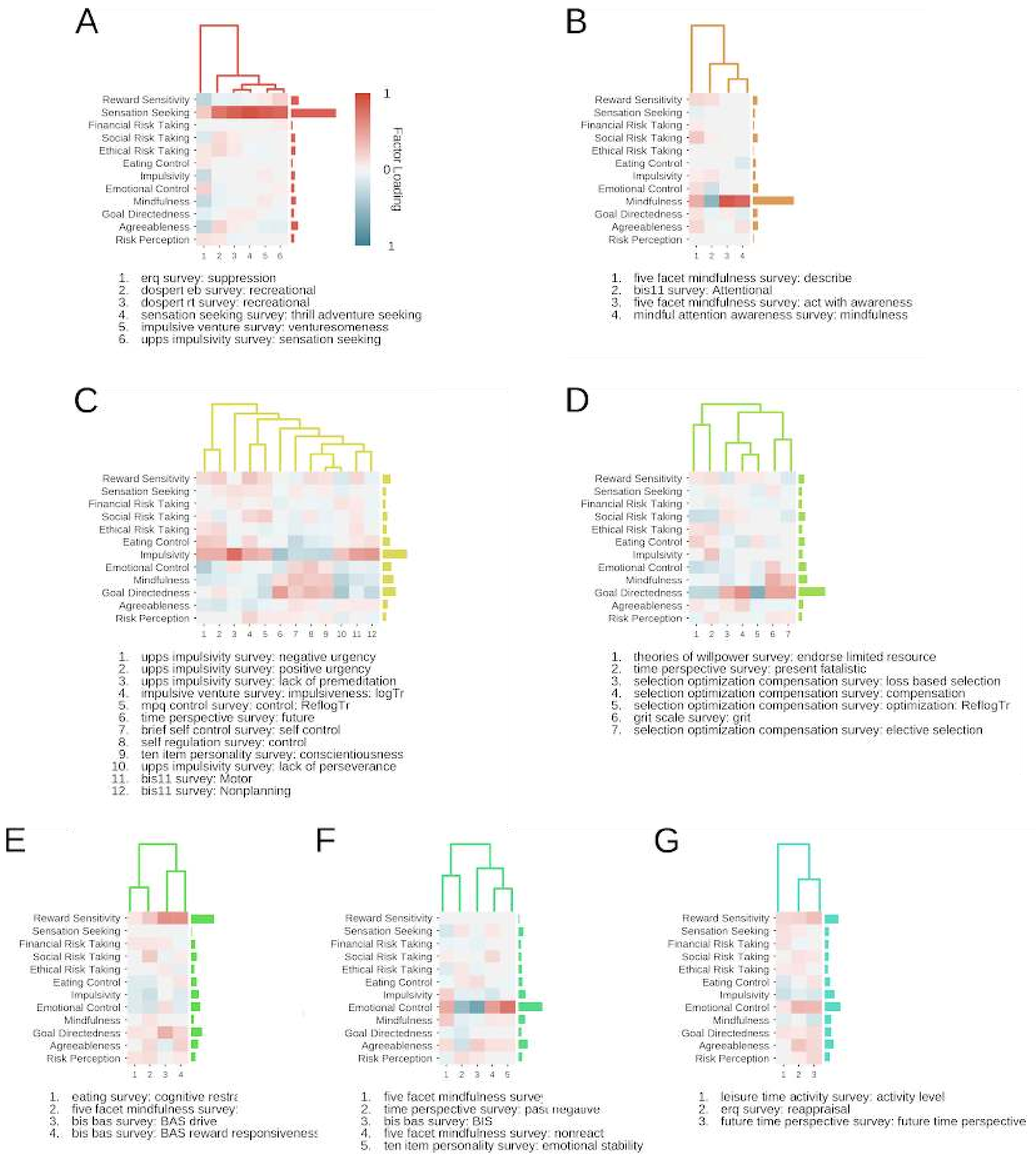


Fig. S12. Clustering quality assessment using silhouette analysis. (A) and (C) depict the silhouette scores for each DV separated by the clustering solution used in the main paper derived from the DynamicTreeCut algorithm. The average silhouette score for these solutions is depicted using a dashed red line. (B) and (D) show the silhouette score using a simpler clustering method - cutting the tree at a single height. The tree was cut at a number of different heights to extract clusters of different sizes (the maximum number of clusters analyzed was a third of the total number of DVs. The silhouette score from the dynamic tree cut solution is shown as a red circle. The dynamic tree cut solution was also used after clustering in "participant space" (see Figs S11-12), and the silhouette score for these solutions are shown as black dots.


A
Reward Sensitivity
Sensation Seeking
Financial Risk Taking
Social Risk Taking
Ethical Risk Taking
Eating Control
Impulsivity
Emotional Control
Mindfulness
Goal Directedness
Agreeableness
Risk Perception
1 2 3 4 5 6
1
0
1
Factor Loading
1. erq survey: suppression
2. dospert eb survey: recreational
3. dospert rt survey: recreational
4. sensation seeking survey: thrill adventure seeking
5. impulsive venture survey: venturesomeness
6. upps impulsivity survey: sensation seeking
B
Reward Sensitivity
Sensation Seeking
Financial Risk Taking
Social Risk Taking
Ethical Risk Taking
Eating Control
Impulsivity
Emotional Control
Mindfulness
Goal Directedness
Agreeableness
Risk Perception
1 2 3 4
1. five facet mindfulness survey: describe
2. bis11 survey: Attentional
3. five facet mindfulness survey: act with awareness
4. mindful attention awareness survey: mindfulness
C
Reward Sensitivity
Sensation Seeking
Financial Risk Taking
Social Risk Taking
Ethical Risk Taking
Eating Control
Impulsivity
Emotional Control
Mindfulness
Goal Directedness
Agreeableness
Risk Perception
1 2 3 4 5 6 7 8 9 10 11 12
1. upps impulsivity survey: negative urgency
2. upps impulsivity survey: positive urgency
3. upps impulsivity survey: lack of premeditation
4. impulsive venture survey: impulsiveness: logTr
5. mpq control survey: control: ReflogTr
6. time perspective survey: future
7. brief self control survey: self control
8. self regulation survey: control
9. ten item personality survey: conscientiousness
10. upps impulsivity survey: lack of perseverance
11. bis11 survey: Motor
12. bis11 survey: Nonplanning
D
Reward Sensitivity
Sensation Seeking
Financial Risk Taking
Social Risk Taking
Ethical Risk Taking
Eating Control
Impulsivity
Emotional Control
Mindfulness
Goal Directedness
Agreeableness
Risk Perception
1 2 3 4 5 6 7
1. theories of willpower survey: endorse limited resource
2. time perspective survey: present fatalistic
3. selection optimization compensation survey: loss based selection
4. selection optimization compensation survey: compensation
5. selection optimization compensation survey: optimization: ReflogTr
6. grit scale survey: grit
7. selection optimization compensation survey: elective selection
E
Reward Sensitivity
Sensation Seeking
Financial Risk Taking
Social Risk Taking
Ethical Risk Taking
Eating Control
Impulsivity
Emotional Control
Mindfulness
Goal Directedness
Agreeableness
Risk Perception
1 2 3 4
1. eating survey: cognitive restra
2. five facet mindfulness survey:
3. bis bas survey: BAS drive
4. bis bas survey: BAS reward responsiveness
F
Reward Sensitivity
Sensation Seeking
Financial Risk Taking
Social Risk Taking
Ethical Risk Taking
Eating Control
Impulsivity
Emotional Control
Mindfulness
Goal Directedness
Agreeableness
Risk Perception
1 2 3 4 5
1. five facet mindfulness surve
2. time perspective survey: past negative
3. bis bas survey: BIS
4. five facet mindfulness survey: nonreact
5. ten item personality survey: emotional stability
G
Reward Sensitivity
Sensation Seeking
Financial Risk Taking
Social Risk Taking
Ethical Risk Taking
Eating Control
Impulsivity
Emotional Control
Mindfulness
Goal Directedness
Agreeableness
Risk Perception
1 2 3
1. leisure time activity survey: activity level
2. erq survey: reappraisal
3. future time perspective survey: future time perspective

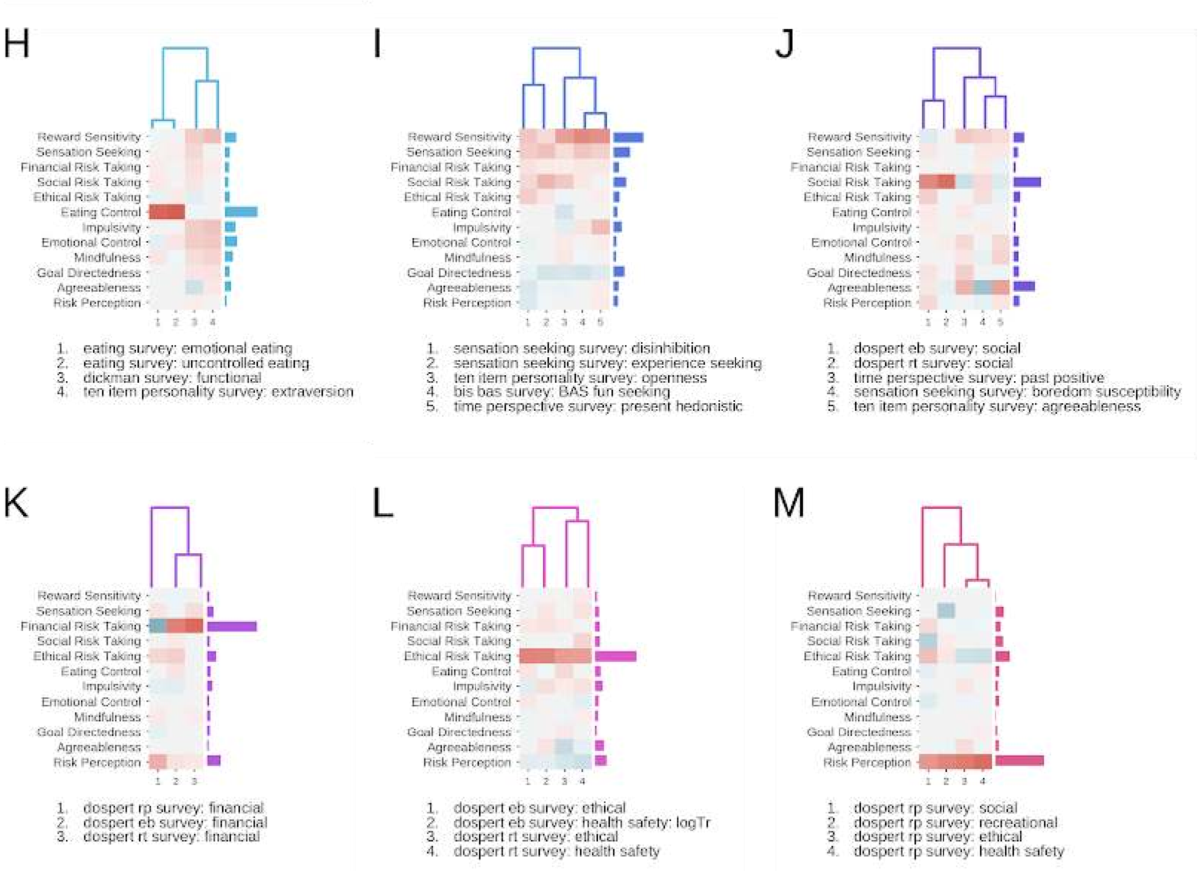
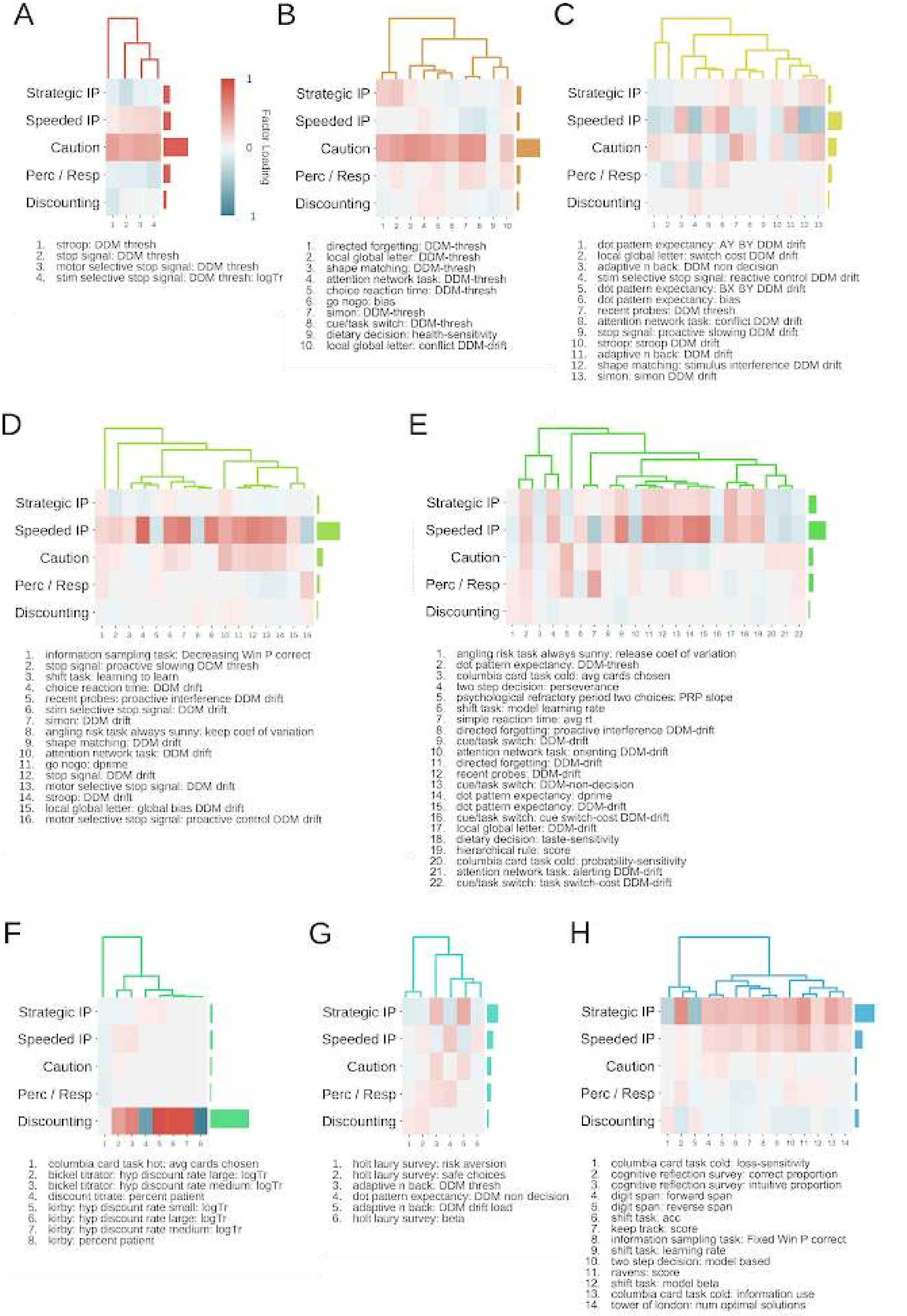


Fig. S13. Survey Clusters. 13 clusters extracted from the full survey ontology dendrogram (Figure 3) using DynamicTreeCut. The clusters are ordered according to the dendrogram from A-M, thus adjacent clusters are more similar to each other.



A

1. stroop: DDM thresh
2. stop signal: DDM thresh
3. motor selective stop signal: DDM thresh
4. stim selective stop signal: DDM thresh: logTr

B

1. directed forgetting: DDM-thresh
2. local global letter: DDM-thresh
3. shape matching: DDM-thresh
4. attention network task: DDM-thresh
5. choice reaction time: DDM-thresh
6. go nogo: bias
7. simon: DDM-thresh
8. cue/task switch: DDM-thresh
9. dietary decision: health-sensitivity
10. local global letter: conflict DDM-drift

C

1. dot pattern expectancy: AY BY DDM drift
2. local global letter: switch cost DDM drift
3. adaptive n back: DDM non decision
4. stim selective stop signal: reactive control DDM drift
5. dot pattern expectancy: BX BY DDM drift
6. dot pattern expectancy: bias
7. recent probes: DDM thresh
8. attention network task: conflict DDM drift
9. stop signal: proactive slowing DDM drift
10. stroop: stroop DDM drift
11. adaptive n back: DDM drift
12. shape matching: stimulus interference DDM drift
13. simon: simon DDM drift

D

1. information sampling task: Decreasing Win P correct
2. stop signal: proactive slowing DDM thresh
3. shift task: learning to learn
4. choice reaction time: DDM drift
5. recent probes: proactive interference DDM drift
6. stim selective stop signal: DDM drift
7. simon: DDM drift
8. angling risk task always sunny: keep coef of variation
9. shape matching: DDM drift
10. attention network task: DDM drift
11. go nogo: dprime
12. stop signal: DDM drift
13. motor selective stop signal: DDM drift
14. stroop: DDM drift
15. local global letter: global bias DDM drift
16. motor selective stop signal: proactive control DDM drift

E

1. angling risk task always sunny: release coef of variation
2. dot pattern expectancy: DDM-thresh
3. columbia card task cold: avg cards chosen
4. two step decision: perseverance
5. psychological refractory period two choices: PRP slope
6. shift task: model learning rate
7. simple reaction time: avg rt
8. directed forgetting: proactive interference DDM-drift
9. cue/task switch: DDM-drift
10. attention network task: orienting DDM-drift
11. directed forgetting: DDM-drift
12. recent probes: DDM-drift
13. cue/task switch: DDM-non-decision
14. dot pattern expectancy: dprime
15. dot pattern expectancy: DDM-drift
16. cue/task switch: cue switch-cost DDM-drift
17. local global letter: DDM-drift
18. dietary decision: taste-sensitivity
19. hierarchical rule: score
20. columbia card task cold: probability-sensitivity
21. attention network task: alerting DDM-drift
22. cue/task switch: task switch-cost DDM-drift

F

1. columbia card task hot: avg cards chosen
2. bickel titrator: hyp discount rate large: logTr
3. bickel titrator: hyp discount rate medium: logTr
4. discount titrate: percent patient
5. kirby: hyp discount rate small: logTr
6. kirby: hyp discount rate large: logTr
7. kirby: hyp discount rate medium: logTr
8. kirby: percent patient

G

1. holt laury survey: risk aversion
2. holt laury survey: safe choices
3. adaptive n back: DDM thresh
4. dot pattern expectancy: DDM non decision
5. adaptive n back: DDM drift load
6. holt laury survey: beta

H

1. columbia card task cold: loss-sensitivity
2. cognitive reflection survey: correct proportion
3. cognitive reflection survey: intuitive proportion
4. digit span: forward span
5. digit span: reverse span
6. shift task: acc
7. keep track: score
8. information sampling task: Fixed Win P correct
9. shift task: learning rate
10. two step decision: model based
11. ravens: score
12. shift task: model beta
13. columbia card task cold: information use
14. tower of london: num optimal solutions

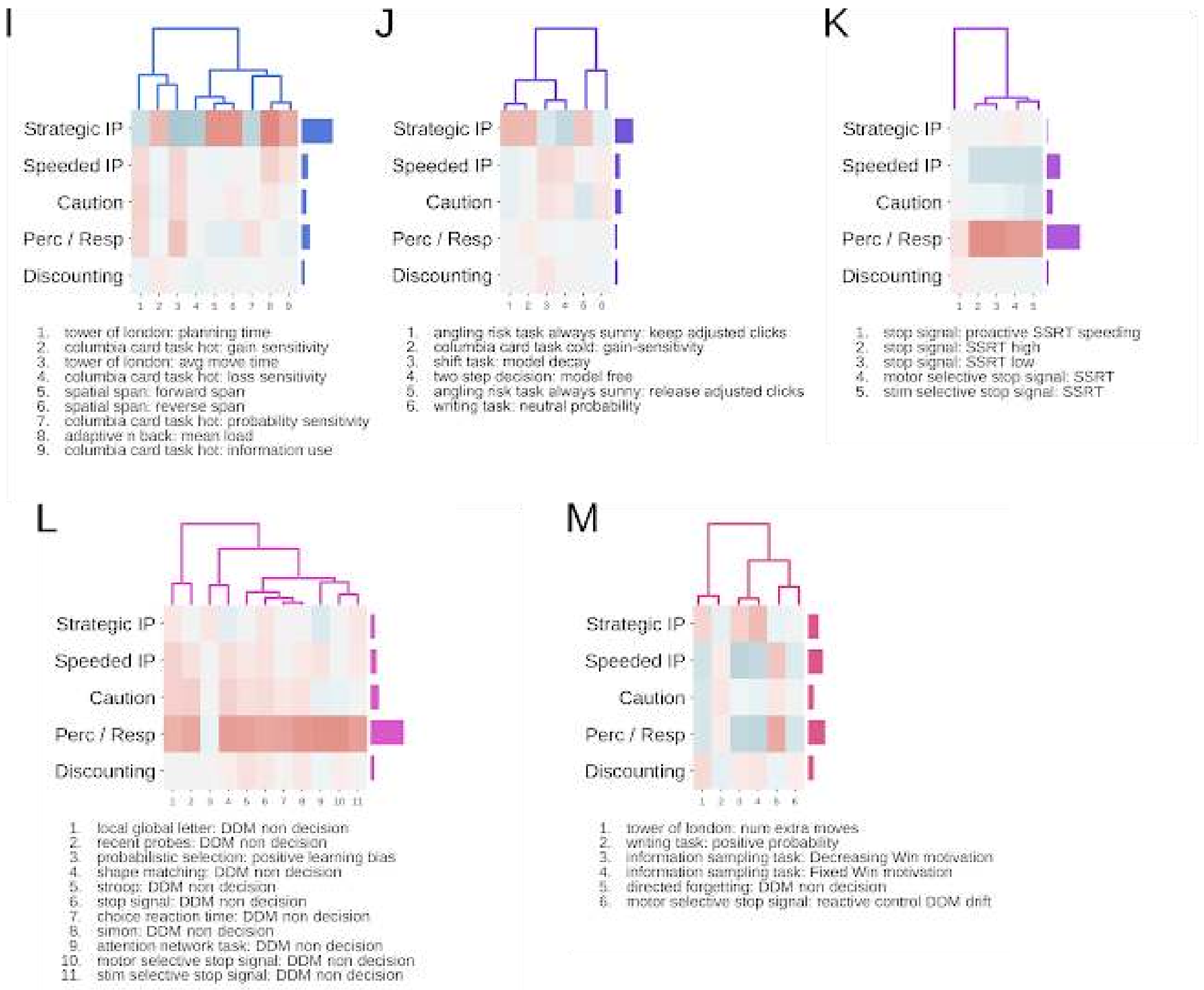


Fig. S14. Task Clusters. 13 clusters extracted from the full task ontology dendrogram (Figure 4) using DynamicTreeCut. The clusters are ordered according to the dendrogram from A-M, thus adjacent clusters are more similar to each other.

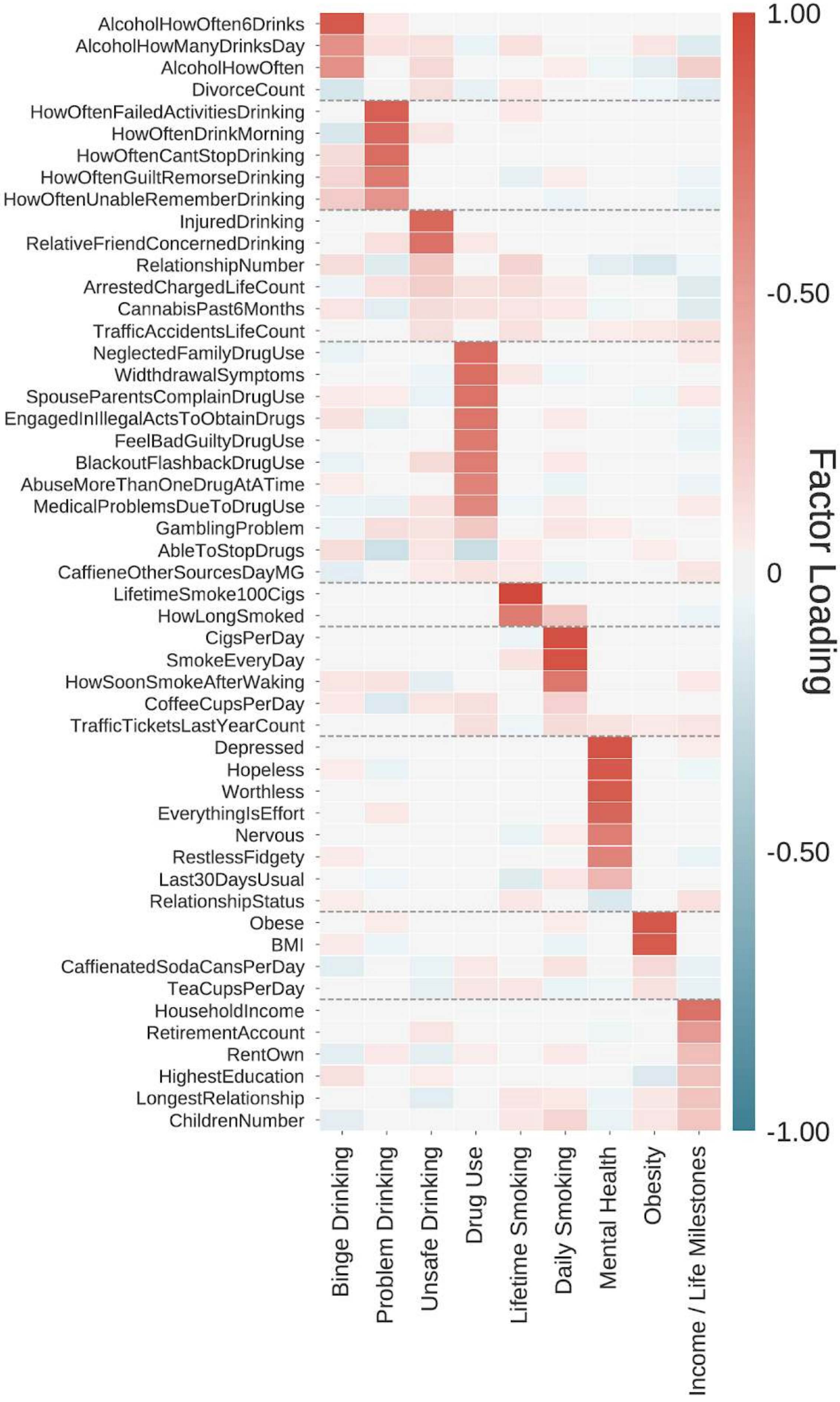

AlcoholHowOften6Drinks
AlcoholHowManyDrinksDay
AlcoholHowOften
DivorceCount
HowOftenFailedActivitiesDrinking
HowOftenDrinkMorning
HowOftenCantStopDrinking
HowOftenGuiltRemorseDrinking
HowOftenUnableRememberDrinking
InjuredDrinking
RelativeFriendConcernedDrinking
RelationshipNumber
ArrestedChargedLifeCount
CannabisPast6Months
TrafficAccidentsLifeCount
NeglectedFamilyDrugUse
WidthdrawalSymptoms
SpouseParentsComplainDrugUse
EngagedInIllegalActsToObtainDrugs
FeelBadGuiltyDrugUse
BlackoutFlashbackDrugUse
AbuseMoreThanOneDrugAtATime
MedicalProblemsDueToDrugUse
GamblingProblem
AbleToStopDrugs
CaffieneOtherSourcesDayMG
LifetimeSmoke100Cigs
HowLongSmoked
CigsPerDay
SmokeEveryDay
HowSoonSmokeAfterWaking
CoffeeCupsPerDay
TrafficTicketsLastYearCount
Depressed
Hopeless
Worthless
EverythingIsEffort
Nervous
RestlessFidgety
Last30DaysUsual
RelationshipStatus
Obese
BMI
CaffienatedSodaCansPerDay
TeaCupsPerDay
HouseholdIncome
RetirementAccount
RentOwn
HighestEducation
LongestRelationship
ChildrenNumber
Binge Drinking
Problem Drinking
Unsafe Drinking
Drug Use
Lifetime Smoking
Daily Smoking
Mental Health
Obesity
Income / Life Milestones
1.00
-0.50
0
-0.50
-1.00
Factor Loading

Fig. S15. Target factor loadings. 9 factors were determined using a BIC criteria for exploratory factor analysis. The 55 target measures are grouped and ordered based on the largest (absolute) factor loading for that target measure. Dotted lines indicate separate groups derived from this criteria, and are used for visualization purposes only.

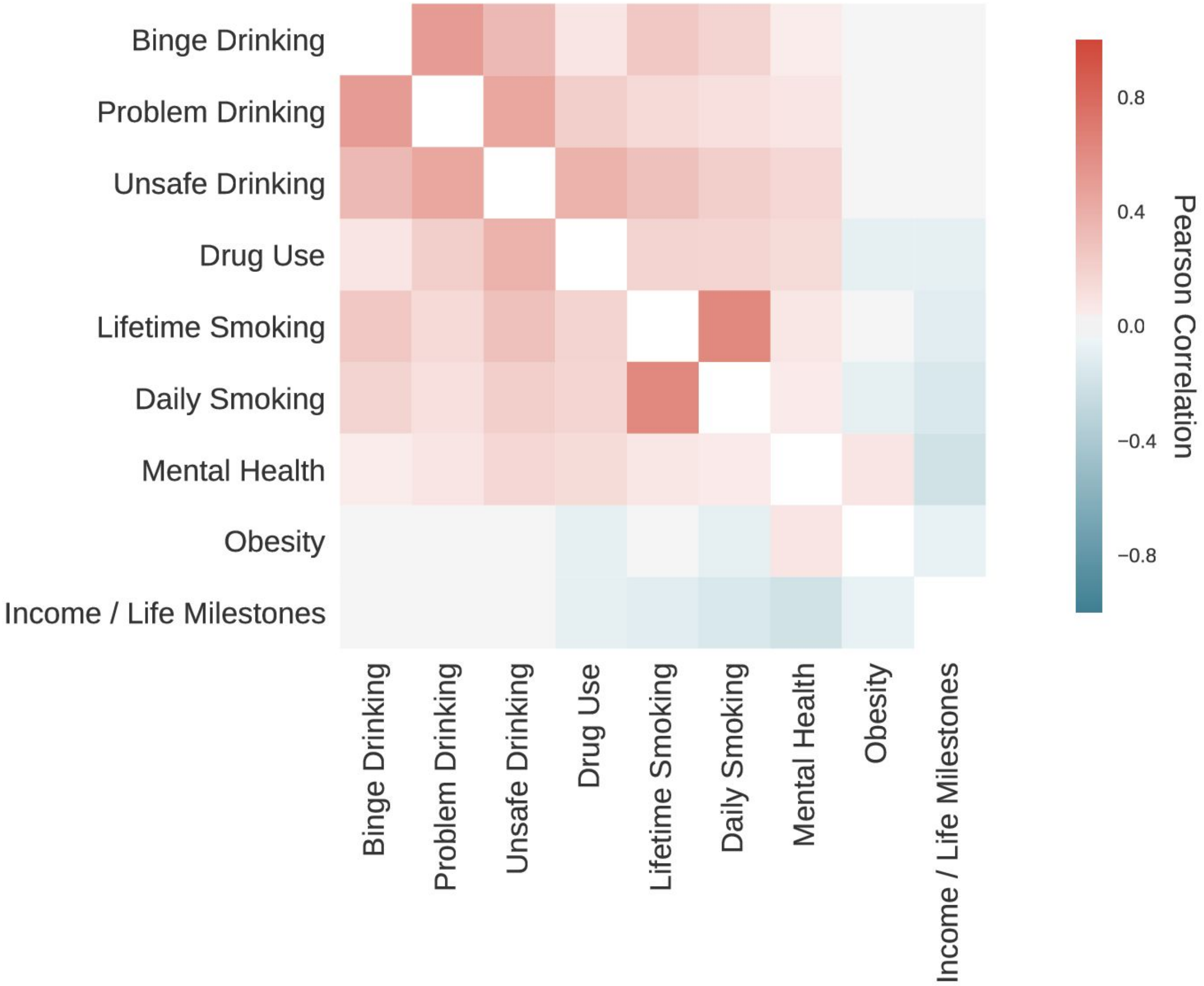


Fig. S16. Target factor correlations. Exploratory factor analysis was performed with oblimin rotation, which allows correlations between factors. The heatmap of these correlations is shown, with the factors ordered by a hierarchical clustering analysis. Note that the colorbar has been scaled based on correlations between factors - *r* values on the diagonal are all 1.

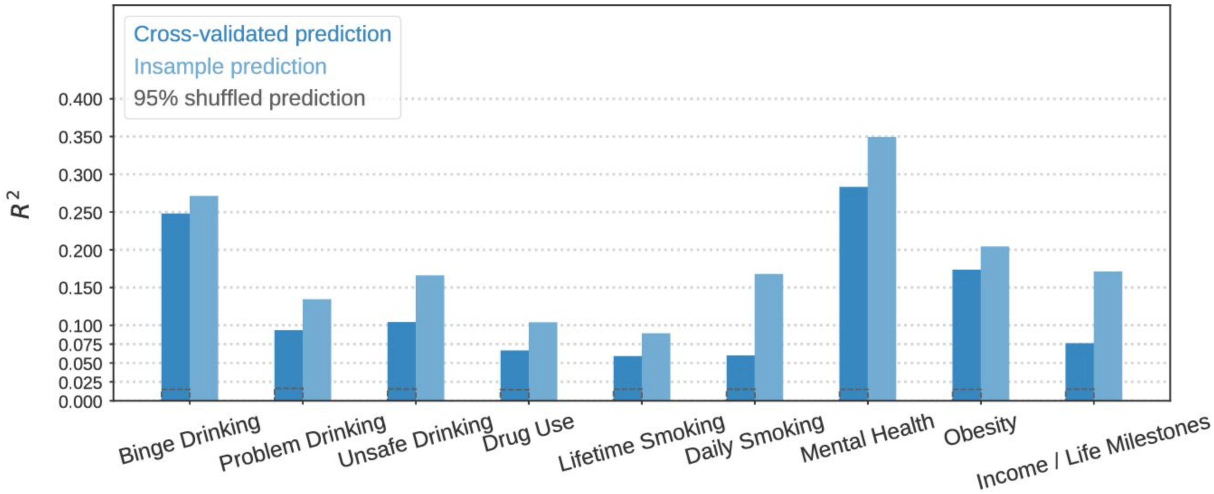


Fig. S17. Prediction of targets using survey DVs. This figure is identical to figure 5 from the main paper, except that the 66 survey DVs were used for prediction instead of the survey factor scores and L1 regularization (lasso) was used instead of L2 (ridge). The prediction performance is qualitatively the same, indicating that the factor solution retained information relevant for prediction.

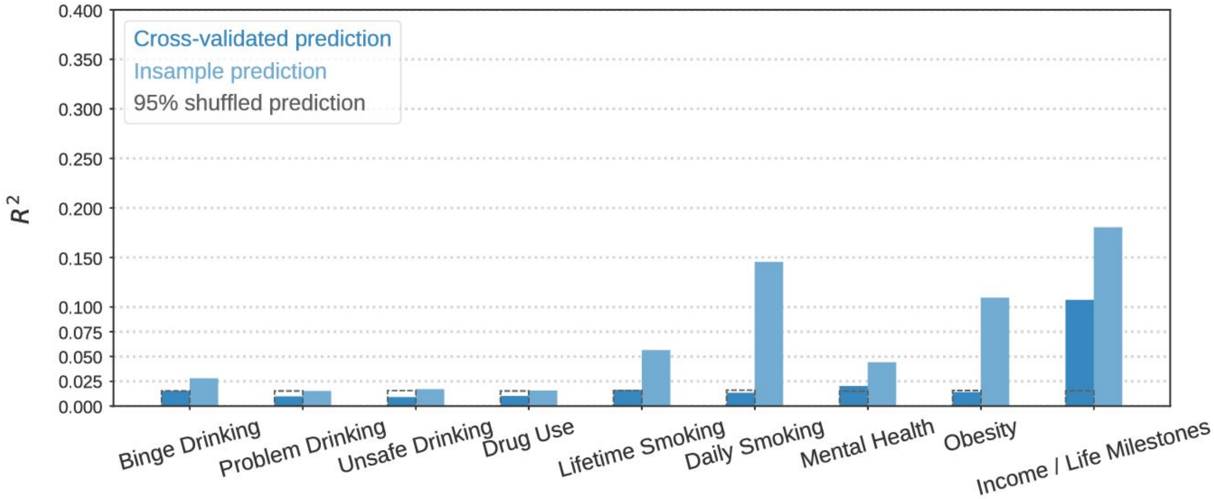


Fig. S18. Prediction of targets using task DVs. This figure is identical to figure 5 from the main paper, except that the 130 task DVs were used for prediction instead of the survey factor scores and L1 regularization (lasso) was used instead of L2 (ridge). The prediction performance is qualitatively the same, indicating that the failure of the tasks to predict the targets was not a function of information loss driven by dimensionality reduction.

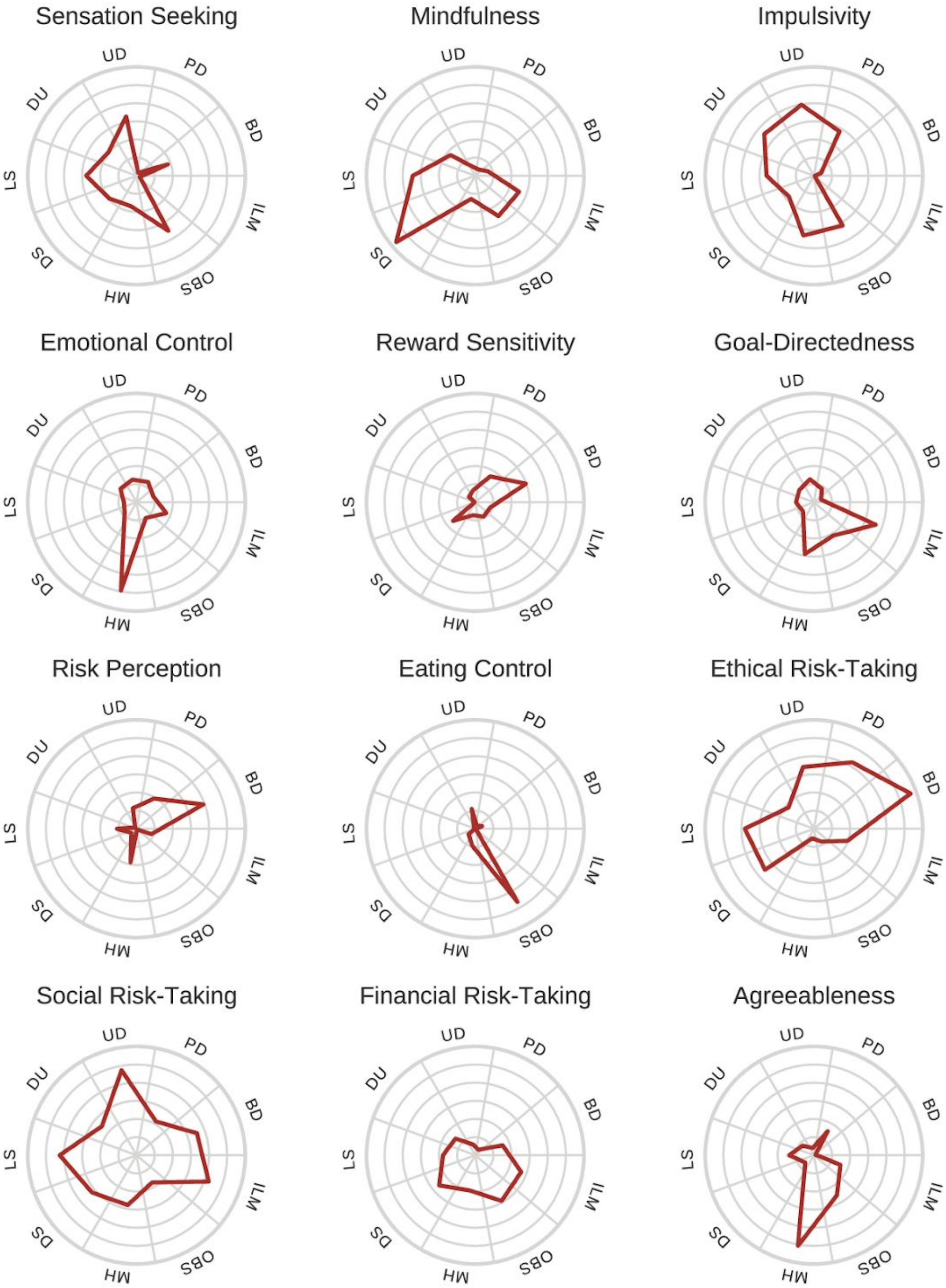


Fig. S19: Survey factor behavioral fingerprints. Each polar plot represents the relationship of each survey factor with the 9 target outcomes. This relationship is defined by the beta-weight

associated with that factor predicting that target outcome. For instance, the factor Eating Control is selectively predictive of obesity, and no other outcome. BD: Binge Drinking, PD: Problem Drinking, UD: Unsafe Drinking, DU: Drug Use, LS: Lifetime Smoking, DS: Daily Smoking, MH: Mental Health, OBS: Obesity, ILM: Income/Life Milestones

References and Notes

45. V. V. Sochat *et al.*, The Experiment Factory: Standardizing Behavioral Experiments. *Front. Psychol.* **7**, 610 (2016).

46. V. A. C. Henmon, The relation of the time of a judgment to its accuracy. *Psychol. Rev.* **18**, 186–201 (1911).

47. R. Ratcliff, R. Childers, *Decisions*, in press.

48. K. Katahira, How hierarchical models improve point estimates of model parameters at the individual level. *J. Math. Psychol.* **73**, 37–58 (2016).

49. D. J. Stekhoven, P. Buhlmann, MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. **28**, 112–118 (2012).

50. F. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

51. S. Epskamp *et al.*, qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Softw.* **48**, 1–18 (2012).

52. M. Bastian, S. Heymann, M. Jacomy, Others, Gephi: an open source software for exploring and manipulating networks. *Icwsm*. **8**, 361–362 (2009).

53. S. Corner, Choosing the right type of rotation in PCA and EFA. *JALT testing & evaluation SIG newsletter*. **13**, 20–25 (2009).

54. J. M. F. ten Berge, W. P. Krijnen, T. Wansbeek, A. Shapiro, Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra Appl.* **289**, 311–318 (1999).

55. W. R. Revelle, psych: Procedures for personality and psychological research (2017) (available at https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research).

56. K. J. Preacher, G. Zhang, C. Kim, G. Mels, Choosing the Optimal Number of Factors in Exploratory Factor Analysis: A Model Selection Perspective. *Multivariate Behav. Res.* **48**, 28–56 (2013).

57. R. Kohavi, Others, in *Ijcai* (Montreal, Canada, 1995), vol. 14, pp. 1137–1145.

58. C. S. Carver, T. L. White, Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *J. Pers. Soc. Psychol.* **67**, 319 (1994).

59. J. H. Patton, M. S. Stanford, E. S. Barratt, Factor structure of the Barratt impulsiveness scale. *J. Clin. Psychol.* **51**, 768–774 (1995).

60. R. M. Roth, P. K. Isquith, G. A. Gioia, *BRIEF-A: Behavior Rating Inventory of Executive Function--adult Version : Professional Manual* (Psychological Assessment Resources, 2005).

61. S. J. Dickman, Functional and dysfunctional impulsivity: personality and cognitive correlates. *J. Pers. Soc. Psychol.* **58**, 95–102 (1990).

62. A.-R. Blais, E. U. Weber, A Domain-Specific Risk-Taking (DOSPERT) Scale for Adult Populations (2006) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1301089).

63. J. J. Gross, O. P. John, Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *J. Pers. Soc. Psychol.* **85**, 348–362 (2003).

64. R. A. Baer, G. T. Smith, J. Hopkins, J. Krietemeyer, L. Toney, Using Self-Report Assessment Methods to Explore Facets of Mindfulness. *Assessment*. **13**, 27–45 (2006).

65. L. L. Carstensen, F. R. Lang, Future time perspective scale. *Unpublished manuscript, Stanford University* (1996).

66. A. L. Duckworth, P. D. Quinn, Development and Validation of the Short Grit Scale (Grit–S). *J. Pers. Assess.* **91**, 166–174 (2009).

67. S. B. G. Eysenck, P. R. Pearson, G. Easting, J. F. Allsopp, Age norms for impulsiveness, venturesomeness and empathy in adults. *Pers. Individ. Dif.* **6**, 613–619 (1985).

68. K. W. Brown, R. M. Ryan, The benefits of being present: mindfulness and its role in psychological well-being. *J. Pers. Soc. Psychol.* **84**, 822 (2003).

69. C. J. Patrick, J. J. Curtin, A. Tellegen, Development and validation of a brief form of the Multidimensional Personality Questionnaire. *Psychol. Assess.* **14**, 150–163 (2002).

70. S. P. Whiteside, D. R. Lynam, The Five Factor Model and impulsivity: using a structural model of personality to understand impulsivity. *Pers. Individ. Dif.* **30**, 669–689 (2001/3).

71. M. Zuckerman, The sensation seeking scale V (SSS-V): Still reliable and valid. *Pers.*

Individ. Dif. **43**, 1303–1305 (2007).

72. M. Zuckerman, Dimensions of sensation seeking. *J. Consult. Clin. Psychol.* **36**, 45 (1971).

73. P. B. Baltes, M. M. Baltes, A. M. Freund, F. R. Lang, The measure of selection, optimization, and compensation (SOC) by self-report. *Max Planck Institute for Human Development, Berlin* (1999).

74. K. B. Carey, D. J. Neal, S. E. Collins, A psychometric analysis of the self-regulation questionnaire. *Addict. Behav.* **29**, 253–260 (2004).

75. M. Kiernan *et al.*, 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**, 1597–1602 (2013).

76. S. D. Gosling, P. J. Rentfrow, W. B. Swann Jr., A very brief measure of the Big-Five personality domains. *J. Res. Pers.* **37**, 504–528 (2003).

77. V. Job, C. S. Dweck, G. M. Walton, Ego Depletion—Is It All in Your Head? *Psychol. Sci.* **21**, 1686–1693 (2010).

78. B. de Lauzon *et al.*, The Three-Factor Eating Questionnaire-R18 is able to distinguish among different eating patterns in a general population. *J. Nutr.* **134**, 2372–2380 (2004).

79. D. R. Lynam, G. T. Smith, S. P. Whiteside, M. A. Cyders, The UPPS-P: Assessing five personality pathways to impulsive behavior. *West Lafayette, IN: Purdue University* (2006).

80. P. G. Zimbardo, J. N. Boyd, in *Time Perspective Theory; Review, Research and Application* (Springer, 2015), pp. 17–55.

81. C. W. Lejuez *et al.*, Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *J. Exp. Psychol. Appl.* **8**, 75–84 (2002).

82. J. Chandler, P. Mueller, G. Paolacci, Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav. Res. Methods*. **46**, 112–130 (2014).

83. M. E. Toplak, R. F. West, K. E. Stanovich, Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Think. Reason.* **20**, 147–168 (2014).

84. C. Primi, K. Morsanyi, F. Chiesi, M. A. Donati, J. Hamilton, The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). *J. Behav. Decis. Mak.* **29**, 453–469 (2016).

85. A. W. MacDonald 3rd *et al.*, A convergent-divergent approach to context processing, general intellectual functioning, and the genetic liability to schizophrenia. *Neuropsychology*.

19, 814–821 (2005).

86. H. E. Rosvold, A. F. Mirsky, I. Sarason, E. D. Bransome Jr, L. H. Beck, A continuous performance test of brain damage. *J. Consult. Psychol.* **20**, 343 (1956).

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

88. D. Badre, A. S. Kayser, M. D'Esposito, Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*. **66**, 315–326 (2010).

89. O. Toubia, E. Johnson, T. Evgeniou, P. Delquié, Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters. *Manage. Sci.* **59**, 613–640 (2013).

90. L. Clark, T. W. Robbins, K. D. Ersche, B. J. Sahakian, Reflection Impulsivity in Current and Former Substance Users. *Biol. Psychiatry*. **60**, 515–522 (2006).

91. H. Pashler, Dual-task interference in simple tasks: data and theory. *Psychol. Bull.* **116**, 220 (1994).

92. J. Raven, Others, in *Handbook of nonverbal assessment* (Springer, 2003), pp. 223–237.

93. A. Radulescu, R. Daniel, Y. Niv, The effects of aging on the interaction between reinforcement learning and attention. *Psychol. Aging*. **31**, 747–757 (2016).

94. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models using lme4. *arXiv [stat.CO]* (2014), (available at http://arxiv.org/abs/1406.5823).

95. A. L. Duckworth, P. D. Quinn, Development and Validation of the Short Grit Scale (Grit – S). *Journal of Personality Assessment*. **91**, 166–174 (2009).

96. P.-O. Harvey *et al.*, Cognitive control and brain resources in major depression: an fMRI study using the n-back task. *Neuroimage*. **26**, 860–869 (2005).

97. S. M. Jaeggi, M. Buschkuehl, J. Jonides, W. J. Perrig, Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*. **105**, 6829–6833 (2008).

98. T. J. Pleskac, Decision making and learning while taking sequential risks. *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 167–185 (2008).

99. J. Fan, B. D. McCandliss, J. Fossella, J. I. Flombaum, M. I. Posner, The activation of attentional networks. *Neuroimage*. **26**, 471–479 (2005).

100. M. N. Koffarnus, W. K. Bickel, A 5-trial adjusting delay discounting task: accurate discount rates in less than one minute. *Exp. Clin. Psychopharmacol.* **22**, 222 (2014).

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

102. T. A. Hare, C. F. Camerer, A. Rangel, Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*. **324**, 646–648 (2009).

103. D. L. Woods *et al.*, Improving digit span assessment of short-term verbal memory. *J. Clin. Exp. Neuropsychol.* **33**, 101–111 (2011).

104. D. E. Nee, J. Jonides, M. G. Berman, Neural Mechanisms of Proactive Interference-Resolution. *NeuroImage*. **38** (2007), pp. 740–751.

105. B. Figner *et al.*, Lateral prefrontal cortex and self-control in intertemporal choice. *Nat. Neurosci.* **13**, 538–539 (2010).

106. A. R. Otto, A. Skatova, S. Madlon-Kay, N. D. Daw, Cognitive Control Predicts Use of Model-based Reinforcement Learning. *J. Cogn. Neurosci.* **27**, 319–333 (2013).

107. C. A. Holt, S. K. Laury, Others, Risk aversion and incentive effects. *Am. Econ. Rev.* **92**, 1644–1655 (2002).

108. A. Miyake *et al.*, The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: a latent variable analysis. *Cogn. Psychol.* **41**, 49–100 (2000).

109. D. B. Yntema, Keeping track of several things at once. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. **5**, 7–17 (1963).

110. K. N. Kirby, N. N. MarakoviĆ, Delay-discounting probabilistic rewards: Rates decrease as amounts increase. *Psychon. Bull. Rev.* **3**, 100–104 (1996).

111. A. R. Aron, T. E. Behrens, S. Smith, M. J. Frank, R. A. Poldrack, Triangulating a Cognitive Control Network Using Diffusion-Weighted Magnetic Resonance Imaging (MRI) and Functional MRI. *Journal of Neuroscience*. **27**, 3743–3752 (2007).

112. M. J. Frank, L. C. Seeberger, R. C. O’Reilly, By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*. **306**, 1940–1943 (2004).

113. C. Stahl *et al.*, Behavioral components of impulsivity. *J. Exp. Psychol. Gen.* **143**, 850–886 (2014).

114. R. C. Wilson, Y. Niv, Inferring Relevance in a Changing World. *Front. Hum. Neurosci.* **5**, 1–14 (2012).

115. C.-H. Lu, R. W. Proctor, The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychon. Bull. Rev.* **2**,

174–207 (1995).

116. P. G. Bissett, G. D. Logan, Selective stopping? Maybe not. *J. Exp. Psychol. Gen.* **143**, 455 (2014).

117. P. G. Bissett, G. D. Logan, Balancing cognitive demands: control adjustments in the stop-signal paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 392–404 (2011).

118. T. Shallice, Specific impairments of planning. *Philosophical Transactions of the Royal Society of London, Biology*. **298**, 199–209 (1982).

119. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron*. **69**, 1204–1215 (2011).