Behavior Research Methods
https://doi.org/10.3758/s13428-020-01534-w

Quantifying mechanisms of cognition with an experiment and modeling ecosystem

Emily R. Weichart[1] · Kevin P. Darby[1] · Adam W. Fenton[1] · Brandon G. Jacques[1] · Ryan P. Kirkpatrick[1] · Brandon M. Turner[2] · Per B. Sederberg[1]

Accepted: 21 December 2020

Abstract

Although there have been major strides toward uncovering the neurobehavioral mechanisms involved in cognitive functions like memory and decision making, methods for measuring behavior and accessing latent processes through computational means remain limited. To this end, we have created SUPREME (Sensing to Understanding and Prediction Realized via an Experiment and Modeling Ecosystem): a toolbox for comprehensive cognitive assessment, provided by a combination of construct-targeted tasks and corresponding computational models. SUPREME includes four tasks, each developed symbiotically with a mechanistic model, which together provide quantified assessments of perception, cognitive control, declarative memory, reward valuation, and frustrative nonreward. In this study, we provide validation analyses for each task using two sessions of data from a cohort of cognitively normal participants ($N = 65$). Measures of test-retest reliability (r: 0.58–0.75), stability of individual differences (ρ: 0.56–0.70), and internal consistency (α: 0.80–0.86) support the validity of our tasks. After fitting the models to data from individual subjects, we demonstrate each model's ability to capture observed patterns of behavioral results across task conditions. Our computational approaches allow us to decompose behavior into cognitively interpretable subprocesses, which we can compare both within and between participants. We discuss potential future applications of SUPREME, including clinical assessments, longitudinal tracking of cognitive functions, and insight into compensatory mechanisms.

Keywords Cognitive assessment · Computational psychiatry · Model-based analysis · Validation

Introduction

In many ways, the development of the research domain criteria (RDoC; Insel et al., 2010) framework by the National Institute of Mental Health has contributed to a fundamental shift in how cognitive disorders are studied and understood. Instead of considering specific disorders in isolation, RDoC is predicated on the idea that cognitive behaviors should be studied at all levels of processing, from neural circuits consisting of multiple brain areas, down to individual neurons. By understanding how patterns of neural activity or connectivity manifest in different cognitive behaviors, the goal of RDoC is to characterize illnesses and injuries in terms of their mechanistic loci in addition to their behavioral symptom profiles. To this end, several studies have used generative modeling techniques to mathematically define the neural processes that underlie distinct patterns of task-related behaviors (Frässle et al., 2018; Friston, Stephan, Montague, & Dolan, 2014; Stephan & Mathys, 2014). Although model-based approaches have been applied to tracking dysfunctions relevant to particular patient groups of interest (Cavanagh et al., 2011; Cockburn & Holroyd, 2010; Frank, Santamaria, O'Reilly, & Willcutt, 2007; Mulder et al., 2010; Wiecki, Poland, & Frank, 2015), the field currently lacks a comprehensive suite of tasks and associated generative models to measure a full range of cognitive mechanisms both within and between participant groups.

Alongside task paradigms designed to target specific cognitive functions, generative models have been used in cognitive psychology for decades to explore and compare mechanistic theories of how neural activity gives rise to behavior. Broadly, generative models fit within a Bayesian framework

✉ Per B. Sederberg
pbs5u@virginia.edu

[1] Department of Psychology, University of Virginia, Charlottesville, VA, USA

[2] Department of Psychology, The Ohio State University, Columbus, OH, USA

Published online: 18 February 2021

consist of (1) a system of equations governing the biological processes relevant to the task; (2) a set of free parameters, whose values are responsible for producing different patterns of data; (3) a likelihood function, representing the probability of the data given a set of parameter values; and (4) a set of prior distributions, which specifies the range of plausible values for each parameter (Frässle et al., 2018; Friston et al., 2014; Stephan & Mathys, 2014). The power of generative modeling is that it allows us to mathematically articulate hypotheses about how different layers of processing function and interact, and in turn, formally test said hypotheses by fitting the models to data (Huys, Maia, & Frank, 2016). Interpreting behavior in terms of latent processes opens a world of possibilities for identifying the networks involved in complex cognitive functions (Herz, Zavala, Bogacz, & Brown, 2016; Nunez, Vandekerckhove, & Srinivasan, 2017). Several reviews have praised generative modeling as a potentially transformative tool for psychiatry, but have also noted the problem of balancing specificity (i.e. capturing relevant deficits in individual patient groups) with generalizability (i.e. being applicable in investigations of other patient groups) when developing tasks and models (Adams, Huys, & Roiser, 2016; Huys et al., 2016; Petzschner, Weber, Gard, & Stephan, 2017).

Here, we present a toolbox of computerized tasks designed to measure an array of cognitive functions, ranging from low-level perceptual decisions to higher-level assessments of risk, and from working memory to long-term associative memory. Our goal was to develop a standard, comprehensive means of assessing cognitive performance, and to provide access to model-based analyses for researchers across areas of expertise. Each task in our toolbox is accompanied by a theory-based generative model designed to quantify the latent processes underlying decisions at the level of each trial. We call our toolbox SUPREME: Sensing to Understanding and Prediction Realized via an Experiment and Modeling Ecosystem. In this article, we aim to accomplish the following. First, we will describe each of the tasks and relate them to specific RDoC cognitive constructs. Second, we will provide the details of our models and the theories about the underlying neural mechanisms that each model represents. Third, we will demonstrate the validity of our models for accurately capturing behavior across a cohort of cognitively normal participants. Finally, we suggest potential extensions of our methods to clinical research in terms of tracking cognitive behavior and mechanistic analogues alongside a diverse array of symptom profiles. Although follow-up work will need to investigate model reliability and applicability to specific clinical diagnoses and other individual difference applications, SUPREME represents a promising step toward a standard method of quantifying and comparing latent cognitive processes both across participant groups and within individuals through time.

Task selection

We aimed to create an ecosystem for comprehensive cognitive assessment, provided by a combination of RDoC construct-targeted tasks and mechanistic computational models. In developing SUPREME, we took care to select tasks that (1) are objective, quantifiable measures of the constructs of interest, (2) can be administered multiple times to the same participants with minimal response biases, (3) are simple enough for participants across a wide range of ages and cognitive abilities to complete, (4) are brief, such that each block takes less than 5 minutes to administer, (5) provide data that are amenable to computational model development and fitting (via constraining task conditions or continuous response time (RT) measures), and (6) span multiple cognitive constructs and different levels of processing complexity. We ultimately selected four tasks that have been mainstays of the cognitive assessment literature for decades, and have been validated by behavioral, neuroimaging, and clinical data to capture group- and condition-level variability in cognitive performance.

Constructs of interest were selected from RDoC, with a primary focus on the *cognitive systems* domain. We selected the random dot motion task (RDM) to measure the construct of *perception*, which encompasses the computations involved in translating sensory input into decision-guiding information. In the task, participants must interpret sensory information in the form of randomly moving dots and identify the direction of most coherent motion (Anstis, 1970; Braddick, 1974). The flanker task (Eriksen & Eriksen, 1974; Kopp, Rist, & Mattler, 1996) was included to measure *attention* and *cognitive control* by requiring participants to make decisions while ignoring task-irrelevant information. The task assesses attention by requiring the purposeful focusing of limited visual processing resources, and assesses cognitive control by requiring the suppression of prepotent response modes. *Declarative memory*—the encoding, storage, and retrieval of events—was measured with the continuous associative binding task (CAB). The CAB task requires participants to remember relationships between paired items, tapping into associative aspects of declarative memory rather than item recognition alone (Gallo, Sullivan, Daffner, Schacter, & Budson, 2004; Henke, Buck, Weber, & Wieser, 1997; Naveh-Benjamin, 2000; Popov, Hristova, & Anders, 2017). Finally, the balloon analogue risk task (BART; Lejuez et al., 2002) was used to assess the *positive* and *negative valence* domains, via the constructs of *reward valuation* and *frustrative nonreward,* respectively. In the task, participants must balance the goal of increasing reward with the risk of loss by interacting with a virtual balloon. Each task was developed alongside a corresponding computational model, which allowed us to gain insight into the mechanistic underpinnings of the cognitive constructs of interest.

Model development

A review by Maia (2015) defined two broad types of model-based analyses in computational psychiatry: (1) *data-driven approaches* in which machine learning methods are used to distinguish among known diagnostic categories of participants, and (2) *theory-driven approaches*, which specify mathematical relationships among variables that contribute to differences in group-level and individual behaviors. Our methods fall into the latter category, such that we developed sets of algorithms to mathematically describe the processes between stimulus onset and response[1]. Free parameters in our models represent cognitively interpretable variables, which, when fit to data, provide quantified measures of latent mechanisms that we could not observe from behavior alone. All of our models operate at the level of the individual trial, making it possible to calculate parameter estimates independently for each participant. Models for each task were developed from a combination of existing decision frameworks (e.g. sequential sampling models) and mathematical articulations of current cognitive theories (e.g. prospect theory). Given that multiple theories exist for how the brain engages certain cognitive processes, our model development procedure was rooted in systematic implementation, fitting, and comparison of contrasting model variants (Kirkpatrick, Turner, & Sederberg, 2019; Weichart, Turner, & Sederberg, 2020). Experiments were designed and models were developed symbiotically in pursuit of the following goals: (1) performance during experiments should involve the cognitive constructs of interest, (2) models should instantiate the cognitive constructs of interest, and (3) experimental conditions should allow for sufficient model constraint, such that the parameter estimates corresponding to the constructs of interest are informative. Here, “informative” refers to parameter estimates that are both accurate and precise, allowing us to meaningfully differentiate whether a mechanism is a plausible component of an individual’s decision-making process. The models presented here provide neurally plausible accounts for the processes underlying the decisions in each task, and also provide the best fits to data compared to alternative accounts as determined by Bayesian comparison analyses.

General methods

Here, we describe procedures for administering our cognitive task battery to a cohort of cognitively healthy participants. Participants completed each task multiple times within a session. Details and illustrations of each task and model are provided in the section entitled *SUPREME cognitive tasks and computational models*.

Participants

Eighty-five participants with a mean age of 20 years (range: 18–43) were recruited from the University of Virginia and the surrounding community via poster advertisement. One to two weeks after completing the first session, a subset of 65 participants returned to the lab and completed a second session. Participants provided written informed consent in accordance with the requirements of the Institutional Review Board at the university. Participants were compensated with $10/hour after each session.

Apparatus

Custom programs using the State Machine Interface Library for Experiments (SMILE; https://github.com/compmem/smile) presented stimuli, tracked timing, and logged responses in all four tasks. Stimuli were presented on a desktop computer equipped with Windows OS connected to a 24-inch, 1920 x 1080-pixel LED display with a refresh rate of 120 Hz. Participants made responses using the outer two keys of a four-key Black Box ToolKit response pad.

Procedure

After providing consent, participants were seated in individual, sound-attenuated rooms and were asked to turn off all electronic devices. An experimenter informed participants that they would be completing four unrelated tasks throughout the experiment, and that they would complete each task multiple times. The experimenter remained in the testing room throughout each session to monitor participants’ engagement. Four blocks each of the RDM, flanker, and CAB tasks and two blocks of BART were presented semi-randomly, such that two blocks of a single task never occurred consecutively. Each block began with instructions and example stimuli to orient participants to the upcoming task while also providing an opportunity to take short breaks as needed. Although tasks differed from one another in objective, all four tasks shared a common two-alternative decision structure. At the end of each block, participants received feedback in the form of a numerical score that was calculated as described in *Validation analyses: Performance metrics*. Across all tasks and blocks, it took participants an average of 40.00 minutes to complete a session (SD = 3.84). Prior to analysis, responses were removed from the RDM and flanker tasks if they occurred faster than 150 ms or slower than 2000 ms post-stimulus onset. Responses were removed from the CAB task if they were faster than 350 ms, but no upper RT criterion was

[1] The model fitting and parameter estimation procedures presented here also provide the opportunity for combining data- and theory-driven approaches, whereby machine learning methods are applied to model parameter estimates rather than summary statistics for subject-level categorization (Huys et al., 2016).

imposed. No RT-based outlier exclusion criteria were applied to data from the BART. While fewer than 1% of trials were excluded from our healthy-subject dataset using these criteria, the SUPREME user is encouraged to carefully consider the outlier specifications that will be appropriate for their participant group of interest.

Model-fitting procedures

As mentioned previously, the overarching goal of our model-based approach was to quantify the latent mechanisms underlying each participant's behavior. Building upon existing models relevant to each task, we mathematically defined the hypothesized neural processes that occur between stimulus onset and response. Models were fit to data from each subject independently using custom programs implemented in RunDEMC (http://github.com/compmem/RunDEMC). While the details of each model are described in the section to follow, a general set of model-fitting procedures are provided here. We first specified the system of equations relevant to each model. Starting parameter values were randomly drawn from a set of prior distributions, which were determined through a series of pilot investigations. Next, the relevant likelihood functions were calculated. For the CAB task, the likelihood function could be calculated directly using analytical solutions described by Navarro and Fuss (2009). For the BART, the probability of generating each choice was determined via a softmax decision rule. Because the sequential dependencies from the presence of leak and lateral inhibition in the models of the flanker and RDM tasks precluded our ability to determine likelihood functions analytically, we instead estimated them by means of simulation and probability density approximation (PDA; Turner & Sederberg, 2014). For all models, the posterior density of the sampled parameter set was calculated as a combination of the likelihood function and the relevant set of prior distributions. New parameter sets were proposed via differential evolution and were updated using Markov chain Monte Carlo (DE-MCMC; Ter Braak, 2006; Turner, Sederberg, Brown, & Steyvers, 2013; Turner & Sederberg, 2012). Depending on the model, this procedure was implemented for 1200–2200 iterations within 50–80 chains in order to calculate full posterior distributions for each model and parameter. MCMC specifications for the number of chains, burn-in iterations, and sampling iterations were selected through a series of preliminary investigations and model comparison studies to provide suitable convergence $\left(\widehat{R} < 1.2\right)$ and good mixing as determined by visual inspection. Model fitting details are provided as part of our publicly available SUPREME codebase (http://github.com/compmem/SUPREME), but users are advised that the provided specifications were only tested on young, cognitively healthy subjects. Because cognitively impaired participants will potentially produce slower, noisier, or otherwise atypical patterns of behavior compared to healthy subjects, users are encouraged to adjust the MCMC specifications as needed to obtain consistent convergence prior to interpreting parameter estimates.

SUPREME cognitive tasks and computational models

In each subsection to follow, we begin by providing background information, details of stimuli, and procedures pertaining to each of our tasks. Because the tasks are based on existing, widely used paradigms, we describe how we developed our specific implementations to be suitably amenable to model-fitting and robustness across sessions. We then describe each of the models that were built to accompany the tasks. Task performance will be reported in terms of custom scoring metrics that will be described in detail in *Validation analyses: Performance metrics*.

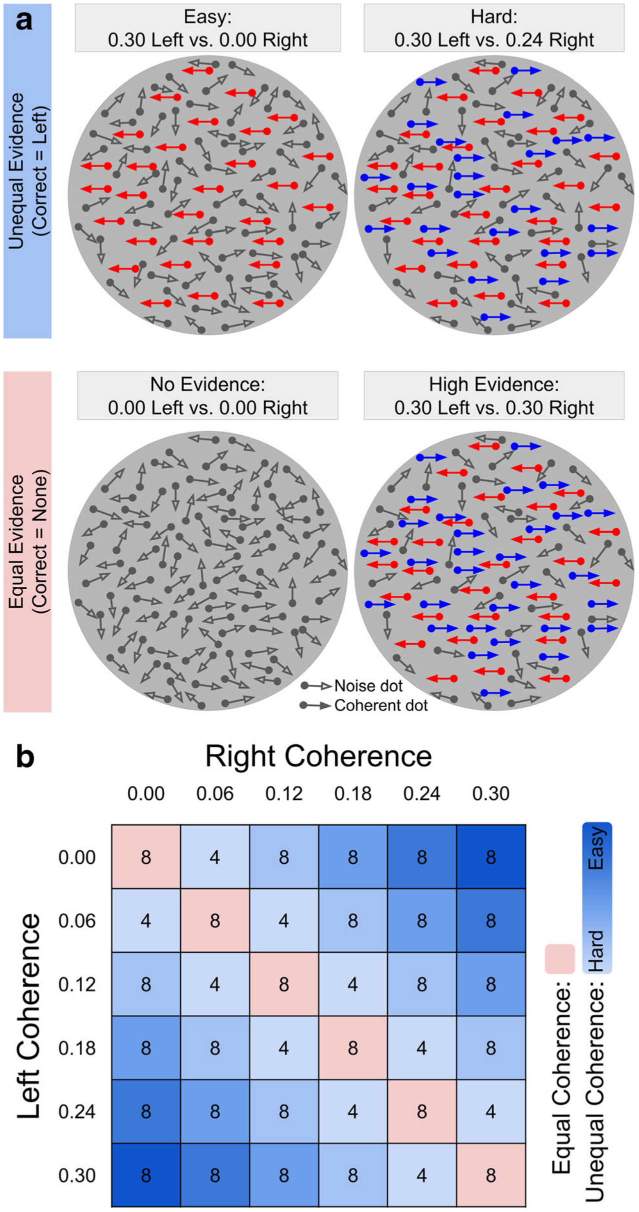
Task 1: Random dot motion

The RDM task is a test of motion detection that has been widely implemented in investigations of perceptual decision-making mechanisms (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff & Starns, 2013; Shadlen & Newsome, 1996; Tsetsos, Gao, McClelland, & Usher, 2012). In the task, participants are asked to indicate the direction of coherent motion amid a cloud of dots that are mostly moving in random directions. Standard behavioral effects are faster, more accurate responses for stimuli with higher proportions of coherently moving dots in one direction relative to the others. In seminal work by Shadlen and Newsome (1996, 2001), the RDM task was used to elucidate the neuron-level representations of motion perception. In particular, the authors found that response accuracy and the firing rates of neurons in the lateral intraparietal (LIP) area both vary as a direct function of dot coherence. The time course of neuronal firing relative to the response has been widely interpreted as evidence of accumulation-to-bound mechanisms in the brain, much like mechanisms described within *sequential sampling models* (SSMs; see Forstmann, Ratcliff, & Wagenmakers, 2016 for review). Broadly, SSMs describe decision-making as the noisy accumulation of evidence for a choice option until a predetermined decision boundary is reached. Several models exist within the SSM framework, each representing subtly different hypotheses about the mechanisms underlying perceptual choice (Brown & Heathcote, 2008; Ratcliff, 1978; Shadlen & Newsome, 2001; Usher & McClelland, 2001). In creating a variant of the RDM task, our goal was to include task conditions that would challenge the assumptions of the various

SSMs and ultimately arrive at a way of calculating constrained, accurate representations of subject-level decision mechanisms. Our variant of the task therefore features bi-directional coherence manipulations, including several conditions of equal coherence in both the left and right directions. In a model-comparison study consisting of six SSMs that had all been successfully fit to RDM data in the past, we found that only one model could capture all the features of behavior across equal coherence conditions in our task variant (Kirkpatrick et al., 2019). The outcome of this work was that we identified the model that effectively maps onto latent mechanisms while also designing a highly constraining task.

Stimuli

In each trial, participants were asked to identify the direction of most coherent motion in a dynamic dot stimulus bound by a circular presentation window with a 200-pixel radius. Stimuli were composed of 100 white dots presented on a gray background, where each dot was three square pixels in size and traveled at a velocity of 200 pixels per second. A random dot pattern with continuous motion was used, but predetermined proportions of dots moved coherently in two target directions. Specifically, we manipulated the proportion of dots moving coherently in both the left (180°) and right (0°) directions within the same stimulus. *Left* and *right* coherence each took on one of six proportion values: 0.00, 0.06, 0.12, 0.18, 0.24, or 0.30. The remaining dots moved at the same velocity as the coherent dots, but in a randomly selected direction. All dots spawned at random locations within the presentation window, and remained on the screen for a life span between 250 and 1250 ms. Dots that reached the end of their life span or the limits of the presentation window were removed and automatically replaced so that there was a constant number of dots on the screen at any given time. Including every pairwise combination of coherence proportions for the left and right directions, there were 36 unique stimuli presented in the task. Most stimuli were presented twice per block. Stimuli in which the difference between left and right coherence was nearly imperceptible (difference = 0.06; 10/36 unique stimuli), however, were presented once per block. We made this decision in an effort to reduce the number of difficult trials from the perspective of the participants. In total, each block contained a total of 62 trials presented in a random order. Task conditions are illustrated in Fig. 1.



Left \ Right	0.00	0.06	0.12	0.18	0.24	0.30
0.00	8	4	8	8	8	8
0.06	4	8	4	8	8	8
0.12	8	4	8	4	8	8
0.18	8	8	4	8	4	8
0.24	8	8	8	4	8	4
0.30	8	8	8	8	4	8

Fig. 1 (**a**) Illustration of conditions in the RDM task. Colors are used to highlight the contrasting directionalities of the dots, but the actual task stimuli used white dots presented against a dark gray background. (**b**) Frequencies of each possible combination of left and right coherence conditions within each block

Procedure

Each RDM block was preceded by a screen with two example stimuli and instructions to select the direction of most coherent dot motion. The instructions remained on the screen until the participant pressed any key on the response pad to proceed. During each trial, a fixation cross appeared on the screen for a jittered duration between 250 and 750 ms. The trial stimulus was then presented and remained on the screen until a response was made. Participants responded by pressing the leftmost key on the response pad to indicate that the dots were moving most coherently to the left, or the rightmost key to indicate that the dots were moving most coherently to the right. Feedback was presented for 500 ms immediately after each response. A green check mark or a red "X" indicated correct and incorrect responses, respectively. The message "Too Fast!" appeared in white text if the participant made a response faster than 100 ms from stimulus onset. Among the

36 unique stimuli in the task, six contained equal proportions of coherent motion in the left and right directions. Because there was technically no correct answer on these trials, response keys were coded randomly such that the "left" response was correct on one half of equal-evidence trials and the "right" response was correct on the other half. Participants were not informed that some trials contained equal coherence in both directions. Across four blocks each consisting of 62 trials, participants completed a total of 248 trials. Each block took an average of 2.06 minutes to complete (SD = 0.32).

Model details

The model included in the current investigation was selected after rigorous comparison of various generative models using Bayesian analytical methods (Kirkpatrick et al., 2019). As previously mentioned, all models in our comparison study were built within the *sequential sampling model* (SSM) framework. In the SSM framework, individual decisions are conceptualized as the stochastic accumulation of evidence through time until a response threshold (α) is reached. This general process has been notably supported by single-unit recordings, which demonstrate accumulation-to-bound patterns of neuronal firing (Churchland, Kiani, & Shadlen, 2008; Shadlen & Newsome, 2001). Across SSMs, an RT is equal to the duration of the decision process plus nondecision time (τ), which comprises perceptual and motor processes. We found that the most successful SSM for fitting data in our RDM paradigm was a variant of the *leaky competing accumulator model* (LCA; Usher & McClelland, 2001, 2004), which features a separate accumulator for each possible choice (in this case, a "left" or "right" response). In the LCA model, accumulators mutually suppress one another via lateral inhibition (β), and evidence passively decays through time (κ). These mechanisms were implemented in the original LCA model to reflect observed biological mechanisms in the brain (Abbott, 1991; Amit, Brunel, & Tsodyks, 1994) and proved necessary for capturing behavior using our task paradigm, particularly in the equal-evidence conditions (Kirkpatrick et al., 2019). When fitting models to RDM data, rates of evidence accumulation (drift rates) corresponding to each possible choice are directly related to the coherence of motion in the stimulus. For stimuli with only one direction of coherent motion, high-coherence stimuli are associated with higher drift rates, reflecting a robust speed and accuracy advantage compared to performance on low-coherence stimuli. To reflect the general, positive relationship between drift rate and coherence without making any strong assumptions about the functional form of the association, drift rates were calculated based on their position along a monotonically increasing sigmoidal function. The functional form of the sigmoid was specified by three free parameters, representing height (a), shift (b), and sharpness (c). A list of free parameters and the mechanisms they represent are provided in Table 1, and an illustration of the model is provided in Fig. 2.

Table 1 Summary of RDM model free parameters

Parameter	Description
α	Decision threshold
τ	Nondecision time
β	Lateral inhibition
κ	Passive decay of evidence
a	Height of drift rate function
b	Shift of drift rate function
c	Sharpness of drift rate function

Model fits

To illustrate the model's ability to capture observed patterns of responses, Figs. 3, 4, and 5 show data produced by our participants contrasted with data generated by our model using each participant's best-fitting parameters. To assess model fits, we first calculated *maximum a posteriori* (MAP) estimates for the parameters. We then input the best-fitting parameter values into the model, and generated 10,000 simulated trials in each task condition. Each simulated trial produced a response (correct or incorrect) and an RT. As a complement to quantitative model validation using Bayesian model comparison techniques in a previous investigation (Kirkpatrick et al., 2019), we used qualitative analyses to verify our model's applicability to the current dataset. We first wanted to see if the model was able to predict the observed pattern of slower, less accurate responses for more similar levels of dot coherence in the left and right directions. Figure 3 shows RT distributions for correct and incorrect responses in each condition of coherence difference, averaged across subjects. As expected, the model predicts a gradient from fast, relatively accurate responses when there is a large difference between coherence values, to slow, chance-level responses when there is no difference between coherence values. This pattern is reflected in Fig. 4 as well, in which observed and model-generated performance scores based on speed and accuracy were calculated within each condition of coherence difference: scores systematically decrease as the differences between coherence values decrease. Finally, Fig. 5 shows the correlation between observed and model-generated scores across task conditions. With a Pearson's r correlation value of 0.97, we conclude that our model provides excellent fits to data and is able to predict the expected patterns of data within each condition.

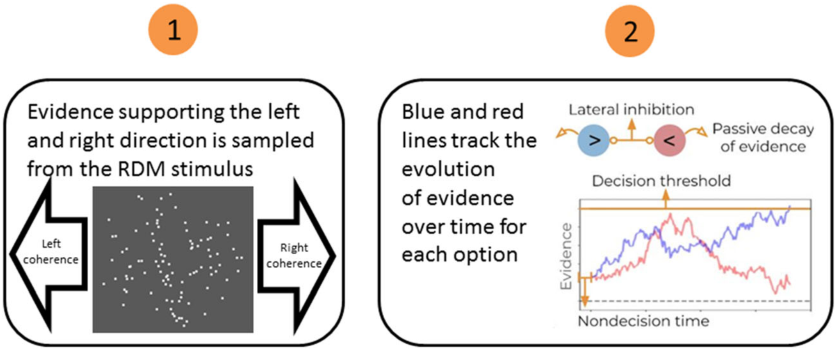

1
Evidence supporting the left and right direction is sampled from the RDM stimulus
Left coherence
Right coherence
2
Blue and red lines track the evolution of evidence over time for each option
Lateral inhibition
Passive decay of evidence
>
<
Decision threshold
Evidence
Nondecision time

Fig. 2 Illustration of the RDM model. (1) Evidence supporting the right and left directions is continuously sampled from the RDM stimulus. (2) The evidence supporting the left direction inhibits the evidence supporting the right direction and vice versa while the evidence supporting both options passively decays over time. The decision is made when the evidence supporting one option is greater than the decision threshold

Task 2: Flanker

The flanker task is a common assessment of inhibitory control, or an individual's ability to ignore goal-irrelevant information and focus on a particular visual input. In the standard paradigm, participants are asked to indicate the direction of a central arrow while ignoring distractors that may be incongruent (<<<><<<) or congruent (>>>>>>>) relative to the target (Eriksen & Eriksen, 1974; Kopp, Rist, & Mattler, 1996). The classic *congruency* effect is that participants are slower and less accurate at responding to incongruent compared to congruent trials, and the magnitude of the effect is often used as a between-subjects index of inhibitory control. In developing our variant of the flanker task, our goal was to create stimuli that were cognitively challenging enough to produce differences in performance among cognitively normal participants while still being simple enough for impaired participants to respond to as well. Our variant differs from the classic paradigm in a number of ways. First, target arrows were presented in the center of a diamond-shaped array consisting of 12 distractor arrows, such that participants were required to inhibit information in both the horizontal and vertical

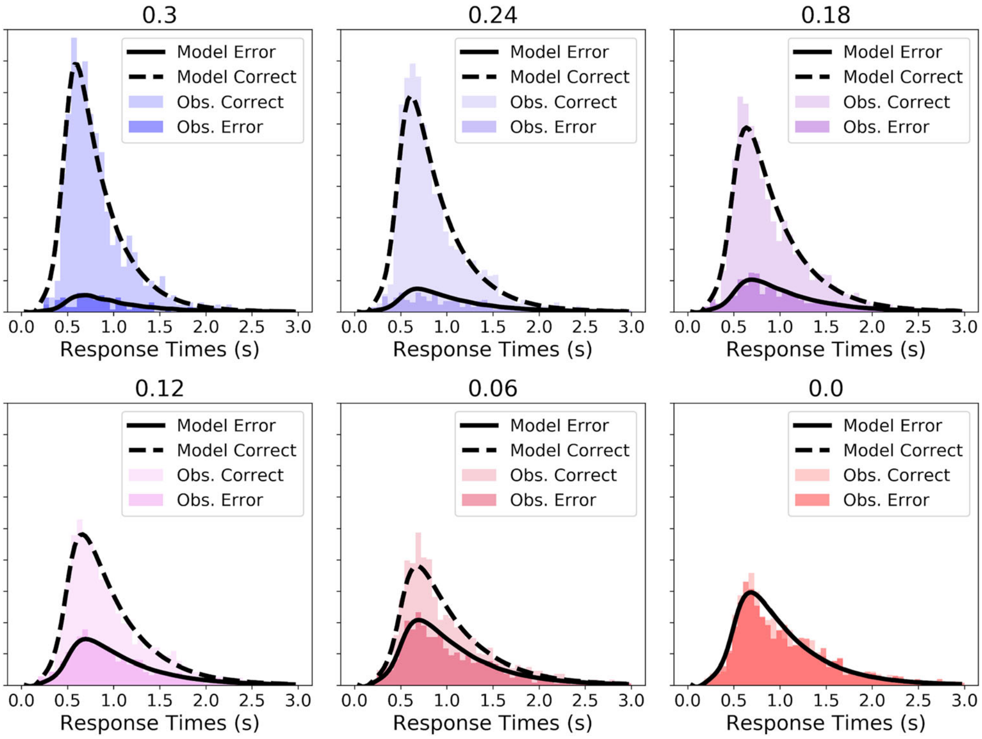


Fig. 3 Observed and model-generated choice-RT distributions in each condition of the RDM task. Observed RT distributions for correct (light-colored histograms) and incorrect (dark-colored histograms) responses were averaged across participants. Models were simulated 10,000 times for each condition, using each participant's best-fitting parameters. Black lines show average model-generated distributions across participants. Facets indicate the absolute difference between left and right coherence

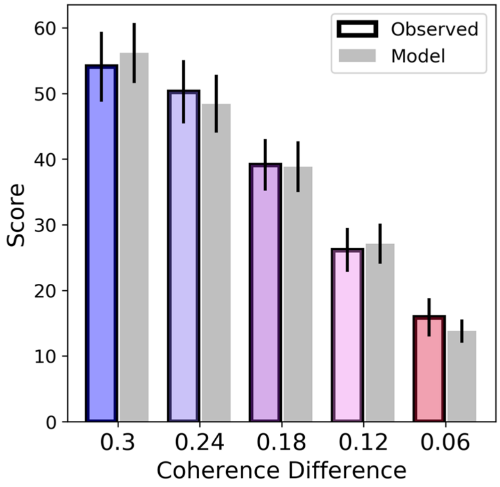


Fig. 4 Observed and model-generated scores in each condition of the RDM task. Models were simulated 10,000 times for each condition using each participant's best-fitting parameters. Scores were calculated based on speed and accuracy (see *Validation analyses: Performance Metrics*. Observed mean and 95% across-subject CIs of scores are shown as colored bars. Mean and 95% across-subject CIs for model-generated scores are shown as gray bars

directions before responding to the target. Second, the configurations of distractor stimuli were based on research demonstrating the *zoom lens* conceptualization of visual attention (Brefczynski & DeYoe, 1999; Tootell et al., 1998). In this framework, attentional resources are oriented around a central point in a graded fashion. To achieve a gradient of difficulty across task conditions, incongruent distractors could therefore be positioned in either the inner or outer layer of the stimulus array. Third, stimuli were presented on the screen at one of eight possible spatial locations on each trial, rather than being presented in the same location throughout the task. Participants were therefore required to dynamically modulate their attention on each trial rather than focusing on a single spatial location throughout the entire task. Examples of stimuli are shown in Fig. 6.

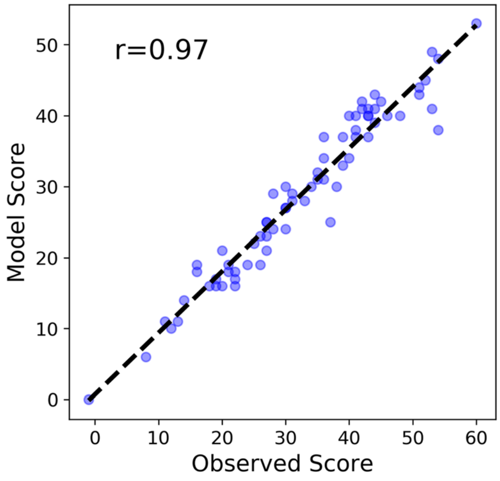


Fig. 5 Correlation between observed and model-generated scores across conditions of the RDM task. The model was simulated 10,000 times for each condition using each participant's best-fitting parameters. Scores were calculated based on speed and accuracy (see *Validation analyses: Performance Metrics*). The line of best fit is shown as a black dashed line

Stimuli

Each stimulus consisted of 13 arrows that were arranged in a diamond formation, and each arrow pointed in the left or right direction. Arrows were 28×28 pixels in size with line widths of two pixels, and each array occupied a 225×225 pixel-sized box on the screen. Participants were instructed to indicate the direction of the arrow in the center of the array while ignoring all distractors. Distractor arrows took on one of three configurations: (1) In the *easy* condition, all 12 distractor arrows pointed in the same direction as the target. (2) In the *moderate* condition, the four inner arrows pointed in the same direction as the target, while the eight outer arrows pointed in the opposite direction. (3) In the *hard* condition, the eight outer arrows pointed in the same direction as the target, while the four inner arrows pointed in the opposite direction. On each trial, the stimulus was presented in one of eight locations around the screen. Possible locations were equidistant from the center of the screen in increments of 45 degrees. Task condition (easy, moderate, hard), target direction (left or right), and screen location (0, 45, 90, 135, 180, 225, 270, or 315 degrees) were counterbalanced and randomized within-block, such that each block consisted of 48 unique, once-presented stimuli.

Procedure

Instructions for the task appeared on the screen, along with examples of the stimuli. Participants completed a 2-minute practice module prior to the first block. In both the practice module and the main task, stimuli were presented in white text on a gray background. Feedback was provided during the practice module to encourage participants to complete the task both quickly and accurately. A green checkmark or a red "X" appeared following correct and incorrect responses respectively, and the message "Too slow!" appeared in white text if participants took longer than 3000 ms to respond. Feedback was provided during practice only, not during the actual task. Participants had the option of completing the practice again before each block, or they were allowed to skip it after completing it once. Regardless, each block began with a summarized instruction screen to orient participants to the upcoming task. The instruction summary remained on the screen until the participant pressed any key on the response pad to proceed. During each trial, a fixation cross appeared in the center of the screen for a jittered duration of 750–1000 ms before being removed. The trial stimulus then appeared on the screen and remained until a response was made. Participants responded by pressing the leftmost key on the response pad

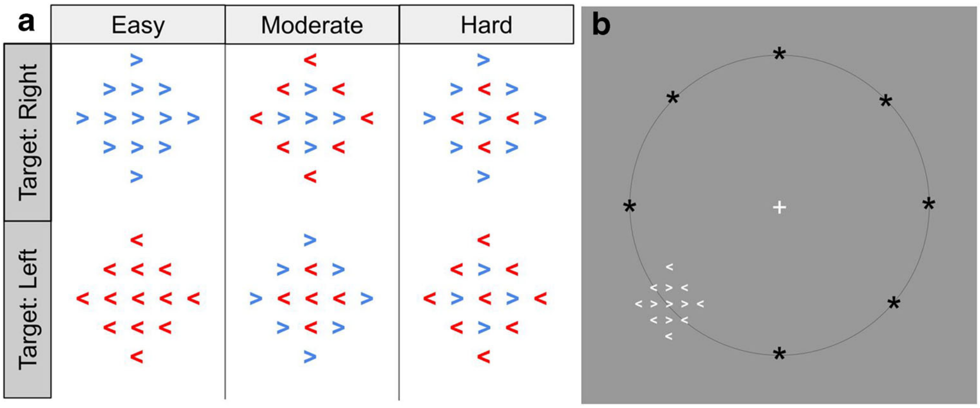


Fig. 6 Illustration of stimuli in each condition of the flanker task (i.e. easy, moderate, and hard). (**a**) Possible stimulus configurations. Colors were used to highlight the contrasting orientations of the arrows, but stimuli in the actual task were presented in white font against a dark gray background. (**b**) Possible stimulus locations. Black asterisks show the eight locations in which a stimulus could have appeared on each trial. The black asterisks and circle are shown here for clarity but were not shown during the task

if the arrow in the center of the array pointed left, and the rightmost key if the center arrow pointed right. Only responses made 150 ms after the stimulus appeared were recorded, and the stimulus was removed from the screen immediately after the participant made a valid response. Participants were given an unlimited amount of time to respond, but were instructed to respond as quickly and accurately as possible. Across four blocks each consisting of 48 trials, participants completed a total of 192 trials. Each block took 1.21 minutes on average to complete (SD = 0.07).

Model details

Similarly to the model of the RDM task, our model of the flanker task was developed within the LCA model framework. As such, the model contains free parameters representing decision threshold (α), nondecision time (τ), lateral inhibition (β), and passive decay of evidence (κ). To calculate the drift rate for each choice, we implemented a variation of time-dependent calculations originally described in the *shrinking spotlight model* (SSP; White, Ratcliff, & Starns, 2011). The SSP draws upon research that suggests that visual attention behaves as a *zoom lens*, such that perceptual resources are allocated around a central target within a finite area that can expand and contract as needed (Brefczynski & DeYoe, 1999; Mesulam, 1990, 1999; Müller, Bartelt, Donner, Villringer, & Brandt, 2003; Tootell et al., 1998). In the model, an attentional spotlight takes the form of a density function for a Gaussian distribution with standard deviation sd_0 centered upon the central target of the flanker task stimulus. Each arrow in the stimulus array occupies one unit of space and has a perceptual input strength of p. Although our task paradigm features distractor items in both the horizontal and vertical directions, we fit our model based only on the arrows along the horizontal midline of the stimuli for the purposes of our current investigation. We made this decision in the interest of reducing computational load after verifying that the one-dimensional spotlight yielded qualitatively similar results to a two-dimensional spotlight as part of a previous investigation that used the same stimuli (Weichart & Sederberg, 2020). In our current variant of the model, the standard deviation of the spotlight shrinks as a function of an endogenous calculation of cognitive control, modified by rate of focus (r_d). Our calculation of cognitive control was based on theoretical descriptions of *reactive control* (Braver, 2012; Braver, Gray, & Burgess, 2008; De Pisapia & Braver, 2006), which suggest that attention is modulated within-trial according to an online calculation of control resources relative to the perceptual conflict within the stimulus. As incongruent stimuli contain evidence for both possible choice options, a greater involvement of attentional mechanisms is required to make a response compared to the case of congruent stimuli. We operationalized the concept of reactive control as a cumulative distance between total evidence and a conflict threshold (δ). Despite neural and behavioral evidence of attention- and cognitive control-mediated changes in the decision process within trials (Czernochowski, 2015; Nigbur, Schneider, Sommer, Dimigen, & Stürmer, 2015; Scherbaum, Fischer, Dshemuchadse, & Goschke, 2011), SSMs typically calculate evidence as a direct function of time. In a recent model comparison study, we showed that our control-based spotlight

Table 2 Summary of flanker model free parameters

Parameter	Description
α	Decision threshold
t_0	Nondecision time
β	Lateral inhibition
κ	Passive decay of evidence
sd_0	Initial spotlight width
p	Perceptual input strength
r_d	Rate of focus
δ	Conflict threshold

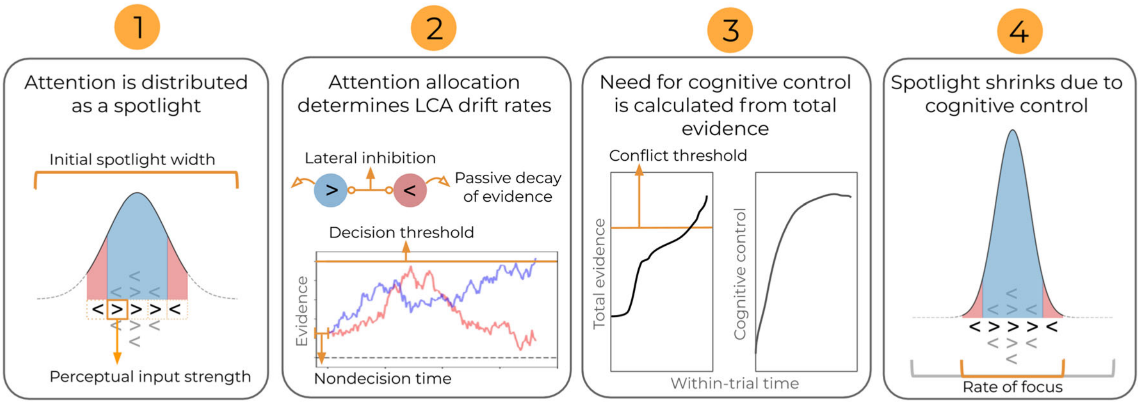

1
Attention is distributed as a spotlight
Initial spotlight width
Perceptual input strength
2
Attention allocation determines LCA drift rates
Lateral inhibition
Passive decay of evidence
Decision threshold
Evidence
Nondecision time
3
Need for cognitive control is calculated from total evidence
Conflict threshold
Total evidence
Cognitive control
Within-trial time
4
Spotlight shrinks due to cognitive control
Rate of focus

Fig. 7 Illustration of the flanker model. (1) Attention, represented as the density function for a Gaussian distribution, is distributed among the arrow stimuli. (2) The area under the attentional spotlight is used to calculate drift rates at each timestep in the leaky competing accumulator (LCA) model. (3) The need for cognitive control is calculated from the distance between the total evidence in the system and a conflict threshold. (4) The attentional spotlight shrinks through time and focuses on the target at a rate governed by cognitive control

implementation within the LCA framework provided better fits to data compared to time-based alternatives, and uniquely mapped onto decision-related signals measured by electroencephalography (EEG; Weichart et al., 2020). A list of free parameters and the mechanisms they represent are provided in Table 2, and Fig. 7 provides a graphical representation of the model's mechanisms.

Model fits

Given that the model was quantitatively validated with Bayesian model comparison techniques in a previous investigation (Weichart et al., 2020), qualitative analyses were applied here to verify that the model was appropriate for the current sample of participants as well. As in the analyses for the RDM task presented previously, we identified a set of MAP parameter estimates. We generated 10,000 trials within each task condition by inputting each set of best-fitting parameters back into the model. Figure 8 shows RT distributions for correct and incorrect responses in each condition, averaged across subjects. The model-generated data reflects the observed pattern of increased proportions of errors and slower responses as we move from the easy to the hard task condition. This pattern of data is shown in Fig. 9 as well, which shows the mean and 95% across-subject confidence intervals of performance scores calculated within-condition. Model performance scores were calculated by generating data within each condition, and applying the same scoring metric that is described in *Validation analyses: Performance metrics* to the simulated choices and RTs. For Fig. 10, we calculated the Pearson's r correlation between each participant's observed score across conditions and the model-predicted scores generated from each participant's best-fitting parameters ($r = 0.82$). Together, these results confirm that our model is providing good fits to data and is able to accurately predict the expected condition-level differences in performance produced by our participants.

Task 3: Continuous associative binding

The CAB task was designed to measure *episodic memory*, which is memory for experiences and the contexts in which

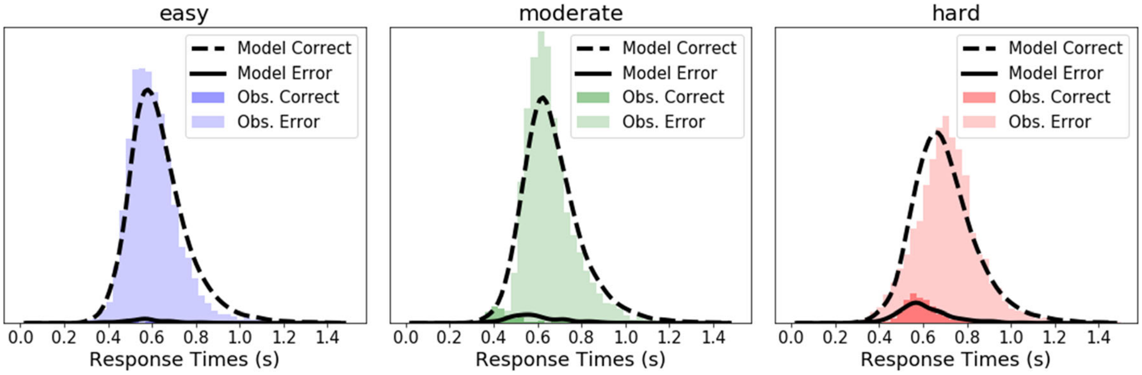


Fig. 8 Observed and model-generated choice-RT distributions in each condition of the flanker task. Observed RT distributions for correct (light-colored histograms) and incorrect (dark-colored histograms) responses were averaged across participants. Models were simulated 10,000 times for each condition, using each participant's best-fitting parameters. Black lines show average model-generated distributions across participants

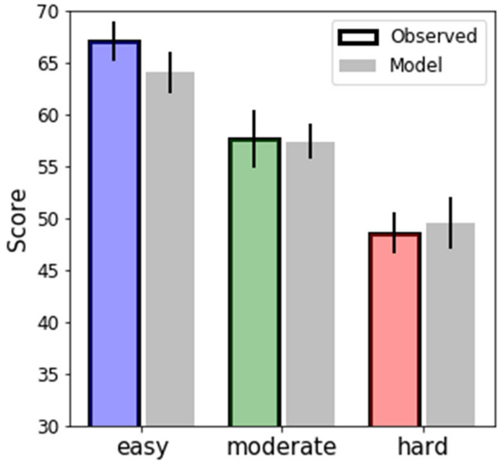


Fig. 9 Observed and model-generated scores in each condition of the flanker task. Models were simulated 10,000 times for each condition using each participant's best-fitting parameters. Scores were calculated based on speed and accuracy as described in Section 4.1. Observed mean and 95% across-subject CIs of scores are shown as colored bars. Mean and 95% across-subject CIs for model-generated scores are shown as gray bars

they occurred (Dickerson & Eichenbaum, 2010; Tulving, 1983; Tulving & Thomson, 1973). Traditional memory tasks, particularly those used in clinical assessments, involve separate phases for *study* and *test* with a delay in between (NIH CB toolbox: Bauer & Zelazo, 2013; Mindstreams: Dwolatzky, 2011; MMSE: Folstein, Folstein, & McHugh, 1975; CAMDEX: Roth et al., 1986). CAB, however, is a variant of a *continuous recognition* task that combines the study and test phases into a single stream of repeating items. For each item in a continuous recognition paradigm, participants respond "old" if they have seen the item before, or "new" if they have not. We selected this format in favor of a traditional study-test paradigm because young children and adults in cognitive decline, due to age or diseases such as mild cognitive impairment (MCI) or Alzheimer's disease (AD), struggle to switch between task phases with different instructions (Gupta, Kar, & Srinivasan, 2009; Hutchison, Balota, & Ducheck, 2010). Our version of the task aims to test *associative* memory in particular, which is the ability to learn relationships between pairs of unrelated items (Anderson & Bower, 1974). Due to its reliance on the hippocampus (Mayes, Montaldi, & Migo, 2007), associative memory exhibits robust age-related changes in childhood (Darby & Sloutsky, 2015a) and older adulthood (Castel & Craik, 2003), as well as due to MCI and AD (Greene, Baddeley, & Hodges, 1996).

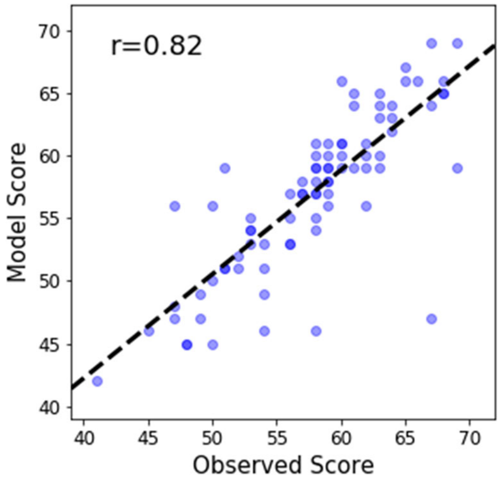


Fig. 10 Correlation between observed and model-generated scores across conditions of the flanker task. The model was simulated 10,000 times for each condition using each participant's best-fitting parameters. Scores were calculated based on speed and accuracy as described in Section 4.1. The line of best fit is shown as a black dashed line

Stimuli

Stimuli were drawn from a database of 2500 full-color images of categorically distinct objects on a white background (Brady, Konkle, Alvarez, & Oliva, 2008). After excluding images that contained people, weapons, or words, a pool of 1995 images remained. A random sample of 96 items was drawn without replacement for each block within a session, and items could not appear in multiple blocks nor in multiple sessions. On each trial, two objects were presented side by side along the horizontal midline of a gray screen. Pairs could be repeated within the block, and participants were instructed to indicate whether they had seen each pair before ("old") or not ("new"). Each item appeared on four trials: in an intact pair presented three times (which we refer to as *intact 1*, *intact 2*, and *intact 3* presentations), and in a *recombined* pair made up of items from different, previously presented intact pairs. Participants were asked to respond "new" to a pair composed of new items, or to a new pairing of recombined (albeit familiar) items; they were asked to respond "old" only when the same pair was repeated exactly. We expected the order of intact 1, intact 2, intact 3, and recombined pairings to affect memory strength, such that a pair presented three times before its component items were recombined would be more strongly remembered than a pair presented only once before being recombined. Prior work has demonstrated that manipulations consistent with this idea have proven a rich source of individual differences between young adults, healthy older adults, and older adults with AD (Gallo et al., 2004; Light, Patterson, Chung, & Healy, 2004), such that younger and more healthy individuals are better able to correctly reject recombined pairs as "new" when the original pairings have been repeated more often. We therefore manipulated whether the intact pairings were presented once, twice, or three times before recombining the items, in the *weak*, *medium*, and *strong* conditions, respectively. We expected that young adults would most easily reject recombined pairs in the strong condition. We also expected intact pairs that were presented following recombined pairs to be remembered less well than

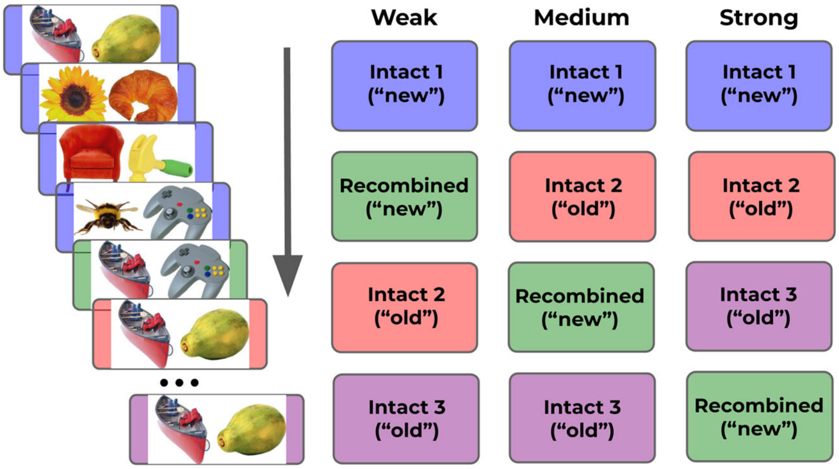


Fig. 11 Illustration of stimuli in the CAB task

pairs that had not been recombined, due to retroactive interference, which occurs when memory for previously learned information is impaired as a result of new learning (Darby & Sloutsky, 2015b). Each condition of memory strength (weak, medium, and strong) and pair type (intact 1, intact 2, intact 3, and recombined) was represented four times within-block, such that each block contained a total of 48 trials. Trials were pseudo-randomized within the confines of the task conditions, and lag (i.e. number of trials) between presentations of each pair was unconstrained. Types of stimuli are illustrated in Fig. 11. Note that although we expect the strength conditions to provide a rich source of individual differences, because our participant sample was limited to young and healthy adults, we collapsed performance across these conditions to simplify our analyses for the current work.

Procedure

Each block was preceded by a 2-minute practice module, in which participants were instructed to respond "new" to new pairs of items, and "old" to pairs they had seen before. Particular emphasis was placed on instructions to respond to the *pair* of items at hand, because individual items could reappear in novel pairings. Items that appeared in the practice module did not appear in the main task. Participants had the option of completing the practice again before each block, or they were allowed to skip it after completing it once. Key mappings were counterbalanced between participants, such that participants with odd subject ID numbers responded with the leftmost key of the response pad to indicate "old" and the rightmost key to indicate "new," while participants with even subject ID numbers responded with the opposite mapping. On each trial, a fixation cross appeared on the screen for a jittered duration of 500–1000 ms. A pair of images appeared, and remained for a fixed duration of 2500 ms. When a response was made, a black rectangle appeared behind the presented objects to provide visual feedback to the participant that their response had been registered. The rectangle remained visible until 2500 ms had elapsed since stimulus onset. Across four blocks each consisting of 48 trials, participants completed a total of 192 trials. Each block took 2.54 minutes on average to complete (SD = 0.02).

Model details

The model of the CAB task is a variant of the *temporal context model* of episodic memory (TCM; Howard, Shankar, Aue, & Criss, 2015; Howard & Kahana, 2002), in which memory is conceptualized as a recency-weighted representation of past experience. In the case of a task in which participants are asked to remember a stream of images, for example, the relevant context for a target item consists of the items that were presented near to the target in time. Memory retrieval involves "jumping back in time" by reinstating contexts bound to previously presented items. In our variant of TCM, a family of decay rates was used to reconstruct past events as described in the *timing from inverse Laplace transform model* (TILT; Shankar & Howard, 2012). The resulting model-generated representation for a given stream of items has a high resolution for what (and when) items occurred in the recent past, and a low resolution for more distant items. The processes described by the TILT model have been supported by single unit "time cell" recordings in the rat hippocampal region (Howard et al., 2014) and the lateral prefrontal cortex of macaque monkeys performing a memory task (Tiganj, Cromer, Roy, Miller, & Howard, 2018). Broadly, our model captures participant responses by estimating the memory strength for each pair of items presented in the task. There are three sources of strength in the model: (1) familiarity of each item within the current context, (2) the overlap between the two items' retrieved contexts, and (3) the mismatch between the two items' retrieved contexts. Familiarity and contextual overlap contribute memory strength in favor of an "old" response, whereas contextual mismatch contributes memory strength in favor of a "new" response. Familiarity is calculated by scanning the temporal context through the past and integrating over all activations that match each object in the target pair (scaled by λ). Estimations of target-relevant activations become less accurate as a function of distance into the past, and noise is drawn from a distribution governed by σ. Contextual overlap and mismatch involve reinstating the previous contexts associated with each item in the target pair, and performing vector calculations to determine the extent to which the contexts do and do not match. Once memory strength is assessed via a combination of familiarity, contextual overlap, and contextual mismatch, the association between the target pair and the current context is updated. The updating process depends on the prediction error associated with observing the target pair within the current context, scaled by a learning rate (α). The context itself is then updated to include the target pair, in consideration of a surprisal signal (modulated by ω) and contextual drift (scaled by δ). Finally, a decision is made by passing the total memory strength of the target pair to a *Wiener first passage of time model* (Stone, 1960), a type of SSM. This decision model features a single noisy accumulator and two opposing boundaries representing each possible choice. Here, the boundary corresponding to a "new" response was set at 0, and the boundary corresponding to an "old" response was represented by a. The starting point (w) could be biased toward either response, or could be situated exactly between the two boundaries. The drift rate was calculated as the difference between the memory strength of the target pair and the strength of the "new" response (ν). The RT was the sum of the decision process and the nondecision time comprised of perceptual and motor processes (t_0). A list of free parameters and the mechanisms they represent are provided in Table 3, and Fig. 12 provides an illustration of the model's mechanisms.

Model fits

Figure 13 illustrates that the model of the CAB task predicts a pattern of performance across conditions that matches the observed behavior of our participants. For these qualitative analyses, we first identified a set of MAP parameter estimates for each participant. We then used each set of best-fitting parameters to generate 10,000 trials in each task condition, and used the performance metric described in *Validation analyses: Performance metrics* to calculate observed and model-predicted performance for each participant and condition. Performance was highest in the "intact 1" condition, which consisted of pairs of previously unseen items. This was considered to be the easiest condition, because participants only had to observe that the pair was "new" without engaging associative memory processes. "Recombined" pairs elicited the worst performance, because each item in the recombined pair had been previously presented within a different pair. Because each item in the pair would have been familiar to the participants, the model predicts a higher likelihood of responding "old" compared to the "intact 1" condition, even though the pairing itself is new. The model and the participants generated higher performance scores in the "intact 3" condition compared to the "intact 2" condition. This is to be expected, given that "intact 3" pairs were previously presented twice and were therefore more familiar to participants than "intact 2" pairs, which were previously presented only once. Figure 14 shows the observed and model-predicted CAB scores for all participants, collapsing across conditions. With a Pearson's r value of 0.93, we have evidence that the model of the CAB task provides excellent fits to data.

Table 3 Summary of CAB model free parameters

Parameter	Description
λ	Scales familiarity strength
σ	Scales context noise
α	Item-context learning rate
ω	Scales context item input
δ	Scales context drift
a	Decision threshold
w	Decision bias
ν	Strength for "new" decision
t_0	Nondecision time

Task 4: Balloon analogue risk

The BART is a test of loss aversion in risky decision-making that was originally developed by Lejuez et al. (2002). On each trial of the task, participants can either collect a small reward, or gamble by inflating a balloon in the hopes of receiving a larger reward in the future. A decision to "pump" the balloon increases the probability that the balloon will explode, resulting in the end of the trial with no reward. An extensive literature has shown that risk-taking behavior on the BART is significantly related to self-reports of risk behaviors (Hopko et al., 2006; Lejuez et al., 2003, 2007) and behavioral correlates of psychopathy (Hunt, Hopko, Bare, Lejuez, & Robinson, 2005). In developing our variant of the task, we implemented subtle changes in an effort to accurately assess loss aversion in a shorter amount of time than in the original paradigm. Changes included more stringent limitations on balloon explosion points, time-varying reward loss to encourage faster responses, and fewer trials (18, compared to 90 in the original paradigm). Through rigorous testing and assessment, our final task provides stable within-subject assessments of risk-taking behavior (as determined by Cronbach's α; see *Validation analyses*) and provides data that suitably constrain posterior parameter estimates of individual-level cognitive mechanisms within our model.

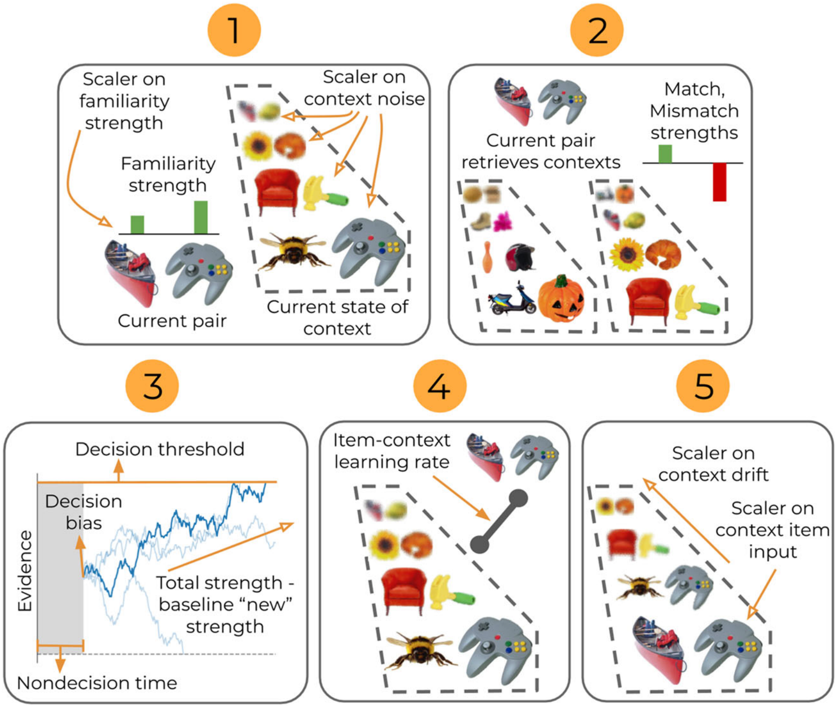

1
Scaler on familiarity strength
Scaler on context noise
Familiarity strength
Current pair
Current state of context
2
Current pair retrieves contexts
Match, Mismatch strengths
3
Decision threshold
Decision bias
Evidence
Total strength - baseline "new" strength
Nondecision time
4
Item-context learning rate
5
Scaler on context drift
Scaler on context item input

Fig. 12 Illustration of the CAB model. 1) Each object pair is evaluated in terms of familiarity strength, based on activation within the current state of context. 2) Previous states of context associated with each item are reinstated and comparedt. 3) Memory strengths are combined to calculate the drift rate in a Wiener-first passage of time decision model. 4) Each item in the presented pair is bound to the current state of context. 5) Context is updated with the presented pair

Stimuli

A simulated image of a balloon connected to a square pump was displayed alongside a rectangle labeled "BANK," as illustrated in Fig. 15. A numerical value in USD was presented within each item in the display (balloon, pump, and bank). The value in the bank represented long-term, permanent earnings over the course of the block (starting value: $1). The value in the balloon represented a short-term, temporary reserve (starting value: $0). The value in the pump represented the amount that *could* be added to the temporary reserve, depending on the participant's decision on the current trial

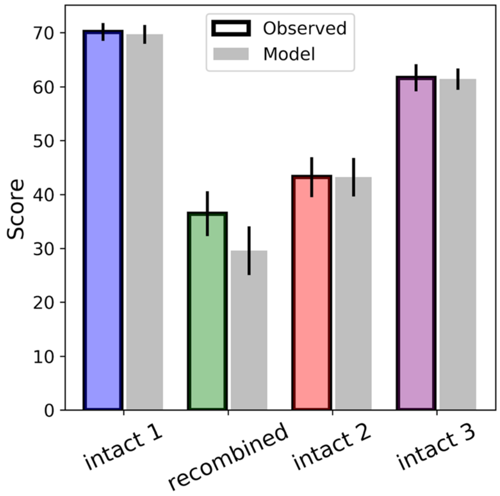


Fig. 13 Observed and model-generated scores in each condition of the CAB task. Models were simulated 10,000 times for each condition using each participant's best-fitting parameters. Scores were calculated based on speed and accuracy (see *Validation analyses: Performance metrics*). Observed mean and 95% across-subject CIs of scores are shown as colored bars. Mean and 95% across-subject CIs for model-generated scores are shown as gray bars

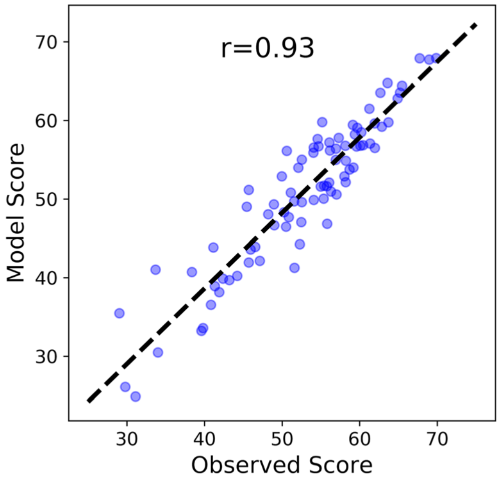


Fig. 14 Correlation between observed and model-generated scores across conditions of the CAB task. The model was simulated 10,000 times for each condition using each participant's best-fitting parameters. Scores were calculated based on speed and accuracy (see *Validation analyses: Performance metrics*). The line of best fit is shown as a black dashed line

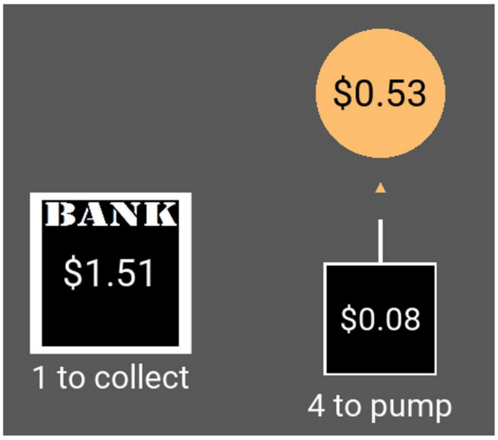


Fig. 15 Illustration of item configurations in the BART

(starting value: $0.05–0.25). On each trial, participants decided whether to "pump" the balloon or to "collect" the earnings in the temporary reserve. When a "pump" response was made, the balloon inflated incrementally, the current pump value was added to the temporary reserve, and a new randomly selected value between $0.05 and 0.25 appeared in the pump. If a balloon exceeded its predetermined explosion point, however, a "pop!" graphic appeared on the screen for 500 ms, the balloon image was removed, and the value in the temporary reserve was lost. Alternatively, a "collect" response transferred the temporary reserve value to the permanent reserve in the bank. In this case, the balloon image moved over top of the bank, the bank value was updated, and the balloon image was then removed. There were three types of balloons, each corresponding to a specific range of explosion points: (1) 1–8 pumps, (2) 8–16 pumps, and (3) 1–16 pumps. The explosion point of each balloon was drawn from a uniform distribution with limits corresponding to the relevant type, and balloons in each type were denoted by a common, randomly selected color within-block. Participants were informed during the instructions that balloon types (specified by color) varied in "quality," but they were not told which colors mapped onto higher or lower quality (i.e. explosion point range). Each block contained six balloons per type, totaling 18 balloons presented in random order.

Procedure

The first block began with a series of instruction screens with still-frame illustrations of the task. The instruction module was self-paced, and participants could proceed to the next screen by pressing any key on the response pad. Subsequent blocks did not contain an instruction module, only a message indicating that the "balloon task" was about to begin. After the instructions, participants completed a 2-minute practice module that was identical to the main task. Practice balloons were gray in color, and had an explosion point range from 1–8 pumps. Participants had the option of completing the practice again before each block, or they were allowed to skip it after completing it once. Key mappings were counterbalanced between participants, such that participants with odd subject ID numbers responded with the leftmost key of the response pad to indicate "pump" and the rightmost key to indicate "collect," while participants with even subject ID numbers responded with the opposite mapping. The appropriate mapping was displayed on the screen throughout the task. Prior to each decision window, a white fixation cross appeared in the center of the pump for a jittered duration of 500–800 ms. The decision window began 100 ms after the fixation cross was replaced with a pump value. Although participants were given an unlimited amount of time to respond, $0.01 was deducted from the bank per 450 ms of the decision period to encourage faster responses. A relevant action sequence began immediately after a response was made. Following a "pop" or "collect" action sequence, a new balloon was attached to the pump after 750 ms with a temporary reserve value of $0. Across two blocks each consisting of 18 trials (balloons), participants completed a total of 36 trials. Each block took 4.05 minutes to complete (SD = 0.90).

Model 4: Balloon analogue risk task

We used the model of the BART originally described by Wallsten, Pleskac, and Lejuez (2005) to capture participant-level sequences of "pump" and "collect" decisions on each trial, incorporating slight modifications to accommodate the features of our task paradigm. The model makes predictions about each decision by considering the number of times the current balloon has already been pumped, as well as previous experience with other balloons of the same type within the task (estimated explosion points: $n_{1,8}$, $n_{1,16}$, $n_{8,16}$). Mechanisms in the model are rooted in *prospect theory* (Kahneman & Tversky, 1979; Tversky & Kahneman, 2000), in which the values ascribed to gains and losses are determined by dissociable, subjective weighting functions (with shapes governed by γ_+ and γ_- for gains and losses, respectively) oriented around a single reference point. This theoretical framework has been supported by a wide array of human neuroimaging, lesion, and neurophysiology work illustrating the neural bases of dichotomous gain and loss functions, and findings that value functions for losses tend to be steeper than those of gains (Schonberg, Fox, & Poldrack, 2011; Trepel, Fox, & Poldrack, 2005). To make a decision whether to "pump" or "collect," the model first calculates the expected value of a "pump" decision based on potential gains and the aversion-weighted potential loss (θ). The probability of a "pump" decision is calculated from the expected value and random variability (β), such that more positive expected values correspond to higher probabilities of "pump" decisions, and more negative expected values correspond to higher probabilities of "collect" decisions. Once a decision is

Table 4 Summary of BART model free parameters

Parameter	Description
$n_{1,8}$	Estimated explosion point (1–8)
$n_{1,16}$	Estimated explosion point (1–16)
$n_{8,16}$	Estimated explosion point (8–16)
γ_+	Shape of reward curve
γ_-	Shape of loss curve
θ	Loss aversion
β	Decision variability
α	Learning rate

made and an outcome is observed, the estimated explosion point of the relevant balloon type is updated as a function of a learning rate (α). A list of free parameters and the mechanisms they represent provided in Table 4, and Fig. 16 shows an illustration of the model.

Model fits

Because the model of the BART accounts for sequential response dependencies within and between trials (balloons), qualitative fits are illustrated for individual subjects in Figs. 17 and 18 instead of collapsing across participants as in our analyses of the other tasks. Figures 17 and 18 show data and model predictions using best-fitting parameters from a highly risk-averse participant (participant s076b session 1 score: 59) and a modestly risk-averse participant (participant s001b session 1 score: 7), respectively. After MAP parameter estimates for each participant were calculated, sequential "pump" and "collect" responses were simulated for each balloon in the order that was observed by the participant. Figures 17 and 18 show distributions of model-predicted stopping points on each balloon, overlaid with the relevant participant's observed behavior. For both subjects illustrated, the model appears to capture both the overall level of exploratory behavior (participant s076b tended to "collect" earlier than the participant s001b) and the variability in stopping behavior (participant s076b was very consistent in stopping behavior across trials, whereas participant s001b showed more variability). To demonstrate model fits across participants, Fig. 19 shows the correlation between observed and model-predicted BART scores using each participant's best-fitting parameters. Consistent with our analyses of the other tasks, model-predicted scores were calculated using the performance metric described in

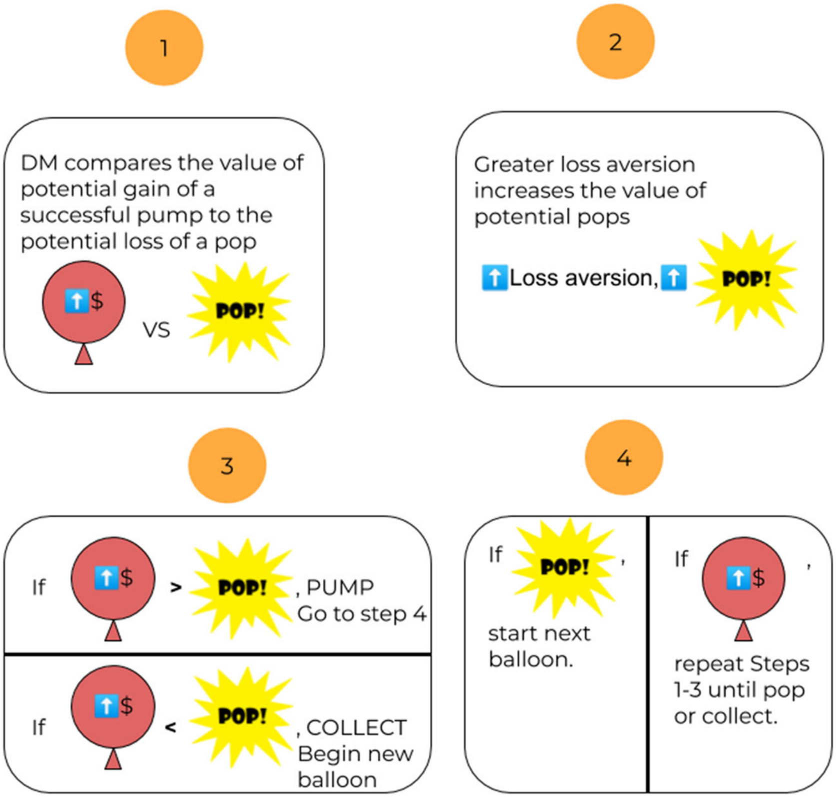

1
DM compares the value of potential gain of a successful pump to the potential loss of a pop
$
VS
POP!
2
Greater loss aversion increases the value of potential pops
Loss aversion,
POP!
3
If
$
>
POP!
, PUMP
Go to step 4
If
$
<
POP!
, COLLECT
Begin new balloon
4
If
POP!
,
start next balloon.
If
$
,
repeat Steps 1-3 until pop or collect.

Fig. 16 Illustration of the BART model. (**1**) The Decision Maker (DM) evaluates the value of pumping [the product of the presented reward and the likelihood of the balloon not popping plus future possible rewards] and compares it to the value of the balloon popping [the product of the total value of the balloon and the likelihood of the balloon popping] at each decision point. (**2**) The DM's loss aversion directly scales the perceived value of the balloon, such that greater loss aversion increases the perceived value of the balloon's total. (3) If the value of pumping is greater than the value of a potential loss, the DM should pump. If the inverse is true, the DM should collect the balloon. (4) If the balloon pops, a new balloon begins

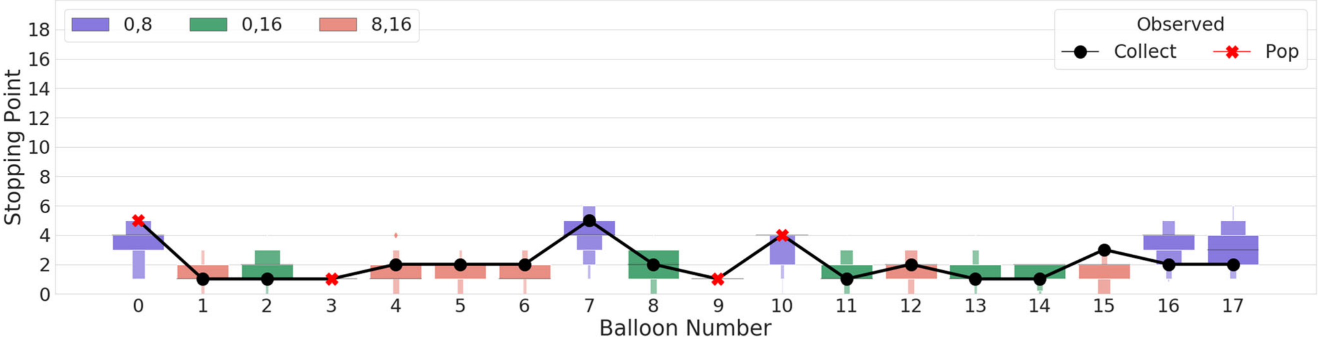


Fig. 17 Observed and model-predicted behavior of a highly risk-averse participant on the BART. Observed behavior of participant s076b in block 2 of session 1 is shown as a black line. Black points indicate the number of "pump" decisions prior to a "collect" response on each balloon. Sequential "pump" and "collect" responses were simulated 1000 times on each balloon using the participant's best-fitting parameters. Boxen plots show the distributions of the model's predicted stopping point on each balloon. Colors indicate the possible explosion range for the balloon at hand

Validation analyses: Performance metrics, and were based on the average stopping point across 10,000 simulations of each balloon. With a Pearson's *r* value of 0.84 between observed and model-predicted scores, we have evidence that the model of the BART provides good fits to data.

Validation analyses

Scoring metrics were developed for each task to assess performance. Validation analyses were applied to each participant's session-level task scores to determine test-retest reliability, stability of individual differences, internal consistency, and construct validity. Before analyzing data, however, binomial tests were applied to assess each participant's engagement with the tasks. In the RDM, flanker, and CAB tasks, the null hypothesis was that participants achieved chance-level accuracy (50% correct) or less in the easiest task condition. The rule for the BART was slightly different, as there was not necessarily a "correct" answer on any given trial. The null hypothesis was that the participant made "collect" responses at least 50% of the time. Across tasks, failure to reject the null hypothesis indicated lack of engagement, and those participants were excluded from further analyses. Participants who did not complete all blocks of a given task were excluded as well. Data from 75 participants in RDM, 84 in flanker, 83 in CAB, and 84 in BART remained from session 1 (out of 85 participants). Data from 62 participants in RDM, 64 in flanker, 65 in CAB, and 64 in BART remained from session 2 (out of 65 participants).

Performance metrics

For each task, we developed a customized metric to calculate participant scores relative to what was considered "optimal" performance. Means and standard deviations of task scores are given in Table 5 to provide a sense of the range of abilities within our cognitively normal, adult sample. For RDM, flanker, and CAB, optimal performance was defined as perfect accuracy across all conditions while maintaining a high response speed. We first calculated an accuracy score by normalizing observed percent correct relative to chance (50%), such that chance accuracy resulted in a score of 0.0 and perfect accuracy resulted in a score of 1.0. We then calculated a speed

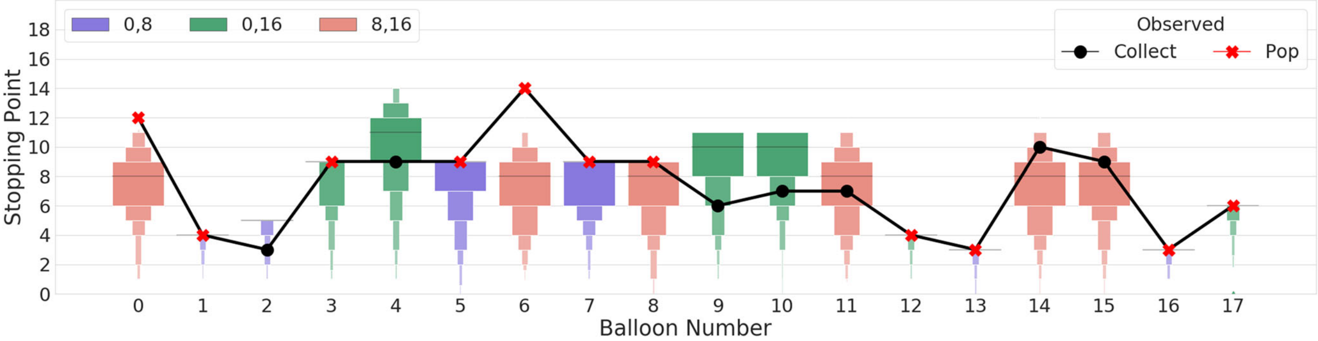


Fig. 18 Observed and model-predicted behavior of a modestly risk-averse participant on the BART. Observed behavior of participant s001b in block 2 of session 1 is shown as a black line. Black points indicate the number of "pump" decisions prior to a "collect" response on each balloon. Sequential "pump" and "collect" responses were simulated 1000 times on each balloon using the participant's best-fitting parameters. Boxen plots show the distributions of the model's predicted stopping point on each balloon. Colors indicate the possible explosion range for the balloon at hand

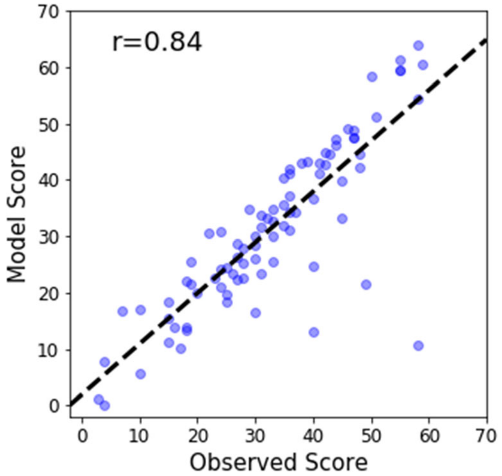


Fig. 19 Correlation between observed and model-generated scores across conditions of the BART. The model was simulated 10,000 times for each observed balloon using each participant's best-fitting parameters. Scores were calculated based on speed and accuracy (see *Validation analyses: Performance metrics*). All scores were greater than zero, indicating that all subjects were risk-averse. The line of best fit is shown as a black dashed line

score by normalizing log RTs within an expected range specific to each task (provided in Table 5), and averaging across trials. An observed average RT equal to the minimum expected RT resulted in a score of 1.0, and an average RT equal to the maximum expected RT resulted in a score of 0.0. The final metric score was calculated by multiplying the accuracy and speed scores together, and converting to a 0 to 100 scale. To earn a high score, participants therefore had to consistently perform quickly and accurately across conditions.

The scoring metric for the BART was necessarily different from that of the other tasks in order to measure both risk-averse and risk-seeking behaviors on a continuous scale. Scores were calculated on a by-balloon basis, then averaged together to produce a final score that could be between −100 and 100. These values represent a distance from "optimal" performance, where negative scores indicate risk-seeking tendencies and positive scores indicate risk-averse tendencies. To calculate a balloon-level score, we first calculated the expected value of a "pump" decision at each possible choice point as the difference between the expected gain if the balloon survives to the next choice point and the expected loss if the balloon pops. Expected gains were based on the probability of the balloon surviving and the amount of money to be added to the temporary store. In contrast, expected losses were based on the probability of the balloon popping given that it had not popped already, and the amount of money in the temporary store that would disappear if the balloon were to pop at the current decision point. Expected *future* value was taken into consideration at each decision point as well. As such, a "pump" decision at the first choice point would be maximally advantageous compared to the other choice points because it accounts for the most future opportunities to increase the reward. Expected values remained positive until the decision point was equal to the median of the current balloon's range of possible explosion points, then became increasingly negative as the potential loss surpassed the potential gain for continuing to make "pump" decisions. Optimal behavior, then, was to "pump" while the expected value was positive, and "collect" when the expected value was near to zero in order to maximize reward. A participant's score on a balloon was equal to the expected value of the participant's final "pump" decision prior to the termination of the trial (i.e. the balloon popped or the participant collected the reward). Values were converted to a scale of −100 to 100 based on the minimum and maximum expected values of the balloon at hand. For the current dataset, all BART scores were above 0, meaning all of our subjects were risk-averse.

Test-retest reliability

Test-retest reliability was assessed with Pearson correlation coefficients between task metric scores in sessions 1 and 2. Test-retest reliability was considered "strong" if r was greater than or equal to 0.80, "moderate" if r was between 0.50 and 0.79, and "weak" if r was less than or equal to 0.50 (Devore & Peck, 1993). All four tasks fell into the "moderate" range (RDM: $r = 0.63$; flanker: $r = 0.73$, CAB: $r = 0.75$, BART: r

Table 5 Scoring metric information and mean performance for each task

	Expected RT range (s)		Task scores: session 1		Task scores: session 2	
	Min.	Max.	Mean	SD	Mean	SD
RDM	0.40	2.00	33.91	11.78	39.91	11.35
Flanker	0.35	1.35	58.05	6.39	60.42	6.62
CAB	0.50	2.50	52.38	11.36	57.78	11.09
BART	–	–	32.39	13.41	31.34	14.22

= 0.58). This is the result we would expect, as our tasks and score metrics were designed to be sensitive to minor fluctuations in cognitive abilities.

Stability of individual differences

Spearman rank order correlations were calculated between task metric scores in sessions 1 and 2 to assess the stability of each participant's scores relative to the rest of the sample. These calculations provide information on the extent to which novelty (session 1) or learning (session 2) may have affected performance (White, Lejuez, & de Wit, 2008). Spearman rank order correlations were moderate across all four tasks (RDM: $\rho = 0.68$, $p < 0.001$; flanker: $\rho = 0.70$, $p < 0.001$; CAB: $\rho = 0.70$, $p < 0.001$; BART: $\rho = 0.56$, $p < 0.001$).

Internal consistency

Cronbach's α provided a measure of internal consistency across all blocks within each session. This measure is particularly important for our purposes, as our task battery was designed for repeated within-session assessments. Values were considered "very high" if greater than or equal to 0.90, "high" if between 0.80 and 0.89, "adequate" if between 0.70 and 0.79, "marginal" if between 0.60 and 0.69, and "low" if less than or equal to 0.59 (Strauss, Sherman, & Spreen, 2006). Internal consistency values were high across all four tasks (RDM: $\alpha = 0.86$; flanker: $\alpha = 0.84$; CAB: $\alpha = 0.80$; BART: $\alpha = 0.84$).

Construct validity

The intercorrelations between task scores are provided in Table 6, and correlations with p values lower than 0.05 are displayed in bold text. For this analysis, we calculated absolute values of BART scores and converted them to a 0 to 100 scale where 100 represented optimal performance and 0 represented maximally risk-seeking or risk-averse behavior. We would expect to observe higher intercorrelations among tasks that were designed to measure a common RDoC domain. Our results show that the tasks relating to the *cognitive systems* domain are significantly intercorrelated (RDM, flanker, and CAB). Intercorrelations involving BART, which represents the *positive* and *negative valence systems* domains, are close to zero. These intercorrelations therefore support the construct validity of our task battery.

Table 6 Intercorrelations among score metrics for tasks included in the cognitive battery. Significant values are presented in bold text

	RDM	Flanker	CAB
RDM	-	-	-
Flanker	**0.25**	-	-
CAB	**0.32**	**0.34**	-
BART	0.08	0.02	0.11

Discussion

The overarching goal of our model-based approach is to quantify the latent mechanisms underlying participant-level behavior. Building upon existing models relevant to each task, we mathematically defined the hypothesized neural processes that occur in between stimulus onset and response. The approach presented here therefore allows for more nuanced analyses of the heterogeneity across participants than is possible with standard behavioral performance metrics alone. This is illustrated in Fig. 20, in which the cognitively normal young adult participants were sorted according to their average overall performance on our battery of cognitive tasks from low to high scores (panel A). Some heterogeneity between participants may be found by breaking overall performance into task-level scores, in that some participants who did badly overall, for example, performed relatively well on one or two individual tasks (panel B). By decomposing task performance into cognitive processes (as estimated by model parameter values, panel C), however, we find a wealth of heterogeneity, in which many individuals who performed similarly on the task have very different profiles of parameter values, which are represented in the figure as colors ranging from red (indicating low values) to blue (indicating high values).

The relationship between brain and behavior is complex. Similar processing modes may lead to different patterns of behavior depending on which brain structures are involved, and the recruitment of different networks could result in nearly identical behavior (Sarter, Berntson, & Cacioppo, 1996). Compensatory mechanisms are a potent example, whereby the brain upregulates alternative processing routes to maintain cognitive performance despite injury, neuropathology, or otherwise limited resources. At one extreme, several studies of traumatic brain injury patients note cognitive reorganization during rehabilitation, such that entire neural architectures reorient to decrease reliance on dysfunctional connections during cognitive operations (see Galetto & Sacco, 2017 for review). Among healthy adults, task-related compensatory increases in anterior neural activity have been noted alongside sensory decline resulting from acute cognitive fatigue (Samuel, Wang, Burke, Kluger, & Ding, 2019) and as a function of age (Cabeza et al., 2004; Madden et al., 2009). Simple individual differences in processing speed at various levels of neural architecture affect how participants solve a task as well, which may or may not result in differences in overt behavioral performance (Schubert, Nunez, Hagemann, & Vandekerckhove, 2019). Analytical methods that are

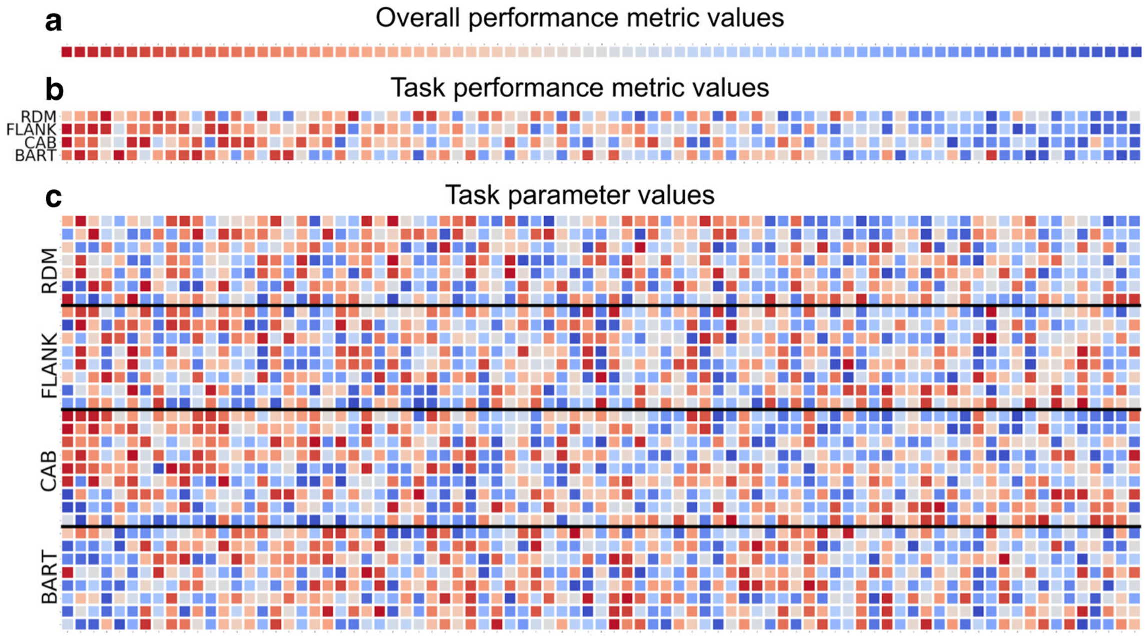


Fig. 20 Visualizing latent heterogeneity in cognitive processes with model parameters. Panel (**a**) presents the overall performance metric score of each participant averaged across tasks, sorted from worst (left) to best (right). Panel (**b**) splits overall performance of the same participants into metric scores for each task, and panel (**c**) presents standardized parameter values for each task in the same group of participants. In all panels, deeper hues of red represent lower values, and deeper hues of blue represent higher values. Each column represents a participant, and each row represents a task (**b**) or a parameter (**c**)

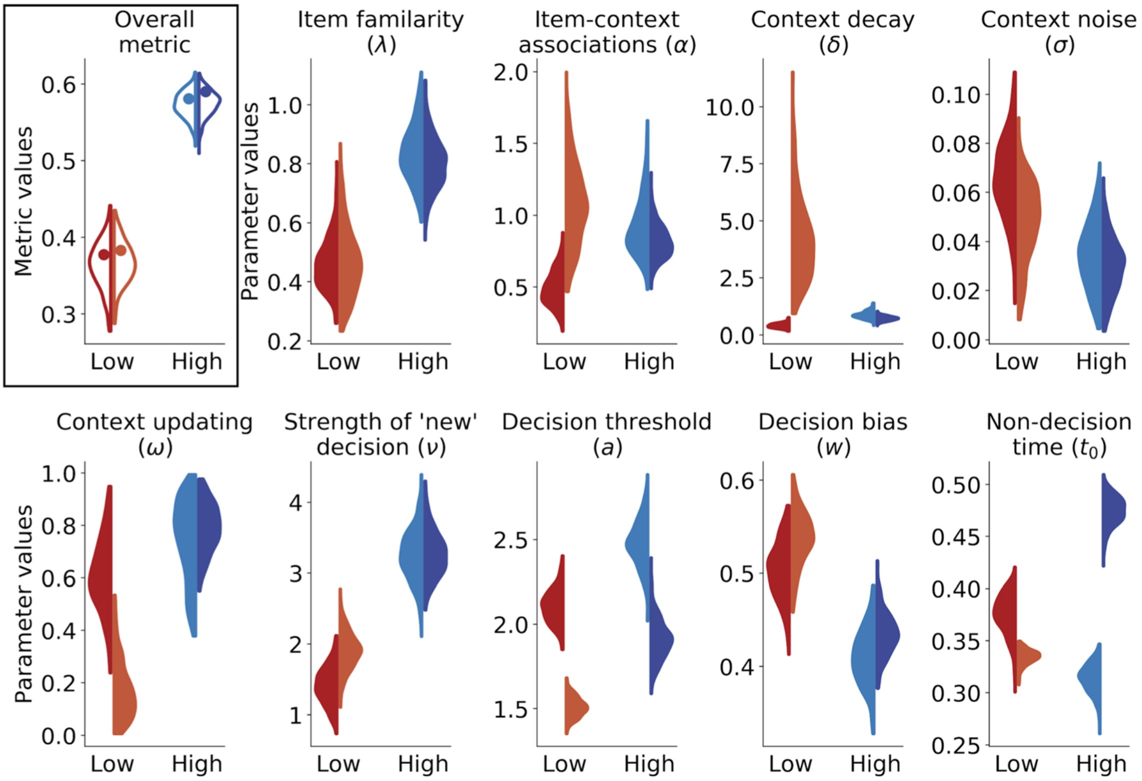


Fig. 21 Four example participants' CAB metric values, posterior predictive distributions (PPDs), and posterior distributions for each model parameter. Two of these participants were low performers, and two were high performers. The top left (boxed) panel shows the observed metric scores as dots for each of these participants. PPDs for these scores are shown as violin plots. The remaining panels show posterior parameter distributions for each of the four participants

considerate of mechanistic heterogeneity across subjects are therefore important for identifying differences in individually-utilized processing routes and for gaining insight into cognitive states. As a proof of concept, studies have shown that machine learning algorithms trained on best-fitting model parameters rather than data-derived summary statistics are more accurate predictors of young versus old age group membership (Wiecki et al., 2015), as well as diagnostic group classification (Petzschner et al., 2017; Weigard, Sathian, & Hampstead, 2020; Wiecki et al., 2016).

Model-based analysis example

One particular strength of our approach is the use of Bayesian analytical procedures, which allow us to identify full distributions of parameter values (the *joint posterior distribution*) that *could* have generated the observed data with some degree of likelihood. As such, we are uniquely positioned to determine both the particular set of values that can optimally recreate a participant's data, as well as a means of quantifying the uncertainty in our measurements of each cognitive construct of interest. Because our models are also *generative* models, meaning they can produce simulated choices and RTs given a set of parameters, we can calculate *posterior predictive distributions* (PPDs) of task scores. PPDs represent the possible ways a participant could have performed, given our model-based assessment of their cognitive state at the time of the test. Despite a participant only completing the task set once and producing a single score, PPDs serve as a model-based confidence interval. With the power to quantify the uncertainty in our measurements via the joint posterior distribution and PPDs, we are able to identify meaningful differences in cognitive acuity at the level of task scores as well as the latent constructs that produced them. This is illustrated in Fig. 21, which presents performance on the CAB task and posterior parameter distributions for four example participants, whom we refer to as participants 1–4, going from left to right. The metric scores for the two low-performing participants were virtually identical, as they were for the two high-performing participants, whereas the scores were quite different between low and high performers. Despite nearly identical performance scores, our model-based analysis detects clear differences in parameter posteriors between the individuals within both the low-performing and high-performing pairs. For example, participant 2 had generally higher values of the α and δ parameters than participant 1, suggesting stronger item-context binding accompanied by faster overall contextual drift. These results demonstrate the explanatory power of our model-based approach, and this type of analysis may be used to glean useful insights into cognitive processes both within individuals across time and between individuals, perhaps with different diagnoses in a case study.

Conclusions

We have developed a battery of cognitive tasks that are simple, objective, quick to administer, and importantly, provide data that are amenable to model development and fitting. Most existing cognitive batteries, which are designed to provide individual metrics that summarize performance, do not provide sufficiently rich data for fitting computational models. In contrast, we developed our tasks and computational models in tandem, defining task conditions that constrain parameter estimates, while capturing fine-grained, individual variation in cognitive abilities. Our tasks, in combination with our set of models, provide a means for quickly and objectively evaluating cognitive mechanisms, beyond what we can learn from behavior alone. Here, we have provided validation analyses and model fits for a cohort of cognitively normal participants. In future work, we will extensively investigate how SUPREME can help answer targeted questions, such as how specific parameters relate to variability in neural signals measured by EEG and fMRI, which parameters are affected by various neuropathologies, and further, how the parameters fluctuate through time according to symptom severity.

Acknowledgements This research was supported by Air Force Research Labs contract FA8650-16-1-6770. The authors do not note any conflicts of interest.

Data Availability The SMILE software package, which was used to code the tasks presented here, and RunDEMC, the software used for all model-fitting, were developed by the corresponding author and are publicly available at http://github.com/compmem. The datasets generated during the current study, experiment code, and model code will additionally be made available at http://github.com/compmem/SUPREME upon publication of the current manuscript.

References

Abbott, L. (1991). Firing rate models for neural populations. In O. Benhar, P. Bosio, P. Giudice, & E. Tabet (Eds.), *Neural networks: From biology to high energy physics* (pp. 179–196). Pisa, Italy: ETS Editrice.

Adams, R., Huys, Q., & Roiser, J. (2016). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery, and Psychiatry*, *87*(1), 53–63.

Amit, D., Brunel, N., & Tsodyks, M. (1994). Correlations of cortical Hebbian reverberations: theory versus experiment. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *14*(11 Pt 1), 6435–6445.

Anderson, J., & Bower, G. (1974). A propositional theory of recognition memory. *Memory & Cognition*, *2*(3), 406–412.

Anstis, S. M. (1970). Phi movement as a subtraction process. *Vision Research*, *10*(12), 1411–1430.

Bauer, P., & Zelazo, P. (2013). IX. NIH Toolbox Cognition Battery (CB): summary, conclusions, and implications for cognitive development. *Monographs of the Society for Research in Child Development*, *78*(4), 133–146.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Braddick, O. (1974). A short-range process in apparent motion. In *Vision Research* (Vol. 14, Issue 7, pp. 519–527). https://doi.org/10.1016/0042-6989(74)90041-8

Brady, T., Konkle, T., Alvarez, G., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Braver, T. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113.

Braver, T., Gray, J., & Burgess, G. (2008). Explaining the Many Varieties of Working Memory Variation: Dual Mechanisms of Cognitive Control. *Variation in Working Memory*, 76–106.

Brefczynski, J., & DeYoe, E. (1999). A physiological correlate of the "spotlight" of visual attention. *Nature Neuroscience*, *2*(4), 370–374.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Cabeza, R., Daselaar, S., Dolcos, F., Prince, S., Budde, M., & Nyberg, L. (2004). Task-independent and task-specific age effects on brain activity during working memory, visual attention and episodic retrieval. *Cerebral Cortex* , *14*(4), 364–375.

Castel, A., & Craik, F. (2003). The effects of aging and divided attention on memory for item and associative information. *Psychology and Aging*, *18*(4), 873–885.

Cavanagh, J., Wiecki, T., Cohen, M., Figueroa, C., Samanta, J., Sherman, S., & Frank, M. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*(11), 1462–1467.

Churchland, A., Kiani, R., & Shadlen, M. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*(6), 693–702.

Cockburn, J., & Holroyd, C. (2010). Focus on the positive: computational simulations implicate asymmetrical reward prediction error signals in childhood attention-deficit/hyperactivity disorder. *Brain Research*, *1365*, 18–34.

Czernochowski, D. (2015). ERPs dissociate proactive and reactive control: evidence from a task-switching paradigm with informative and uninformative cues. *Cognitive, Affective & Behavioral Neuroscience*, *15*(1), 117–131.

Darby, K., & Sloutsky, V. (2015a). The cost of learning: interference effects in memory development. *Journal of Experimental Psychology. General*, *144*(2), 410–431.

Darby, K., & Sloutsky, V. (2015b). When Delays Improve Memory: Stabilizing Memory in Children May Require Time. *Psychological Science*, *26*(12), 1937–1946.

De Pisapia, N., & Braver, T. (2006). A model of dual control mechanisms through anterior cingulate and prefrontal cortex interactions. *Neurocomputing*, *69*(10-12), 1322–1326.

Devore, J., & Peck, R. (1993). *Statistics: Exploration and Analysis*. Arden Shakespeare.

Dickerson, B., & Eichenbaum, H. (2010). The episodic memory system: neurocircuitry and disorders. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 86–104.

Dwolatzky, T. (2011). The mindstreams computerized assessment battery for cognitive impairment and dementia. *The 4th International Conference on Pervasive Technologies Related to Assistive Environments*. https://doi.org/10.1145/2141622.2141681

Eriksen, B., & Eriksen, C. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Folstein, M., Folstein, S., & McHugh, P. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641–666.

Frank, M., Santamaria, A., O'Reilly, R., & Willcutt, E. (2007). Testing Computational Models of Dopamine and Noradrenaline Dysfunction in Attention Deficit/Hyperactivity Disorder. *Neuropsychopharmacology*, *32*(7), 1583–1599.

Frässle, S., Yao, Y., Schöbi, D., Aponte, E., Heinzle, J., & Stephan, K. (2018). Generative models for clinical applications in computational psychiatry. *Wiley Interdisciplinary Reviews. Cognitive Science*, *9*(3), e1460.

Friston, K., Stephan, K., Montague, R., & Dolan, R. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158.

Galetto, V., & Sacco, K. (2017). Neuroplastic Changes Induced by Cognitive Rehabilitation in Traumatic Brain Injury: A Review. *Neurorehabilitation and Neural Repair*, *31*(9), 800–813.

Gallo, D., Sullivan, A., Daffner, K., Schacter, D., & Budson, A. (2004). Associative recognition in Alzheimer's disease: evidence for impaired recall-to-reject. *Neuropsychology*, *18*(3), 556–563.

Greene, J., Baddeley, A., & Hodges, J. (1996). Analysis of the episodic memory deficit in early Alzheimer's disease: evidence from the doors and people test. *Neuropsychologia*, *34*(6), 537–551.

Gupta, R., Kar, B., & Srinivasan, N. (2009). Development of task switching and post-error-slowing in children. *Behavioral and Brain Functions*, *5*, 38.

Henke, K., Buck, A., Weber, B., & Wieser, H. G. (1997). Human hippocampus establishes associations in memory. *Hippocampus*, *7*(3), 249–256.

Herz, D., Zavala, B., Bogacz, R., & Brown, P. (2016). Neural Correlates of Decision Thresholds in the Human Subthalamic Nucleus. *Current Biology*, *26*(7), 916–920.

Hopko, D., Lejuez, C., Daughters, S., Aklin, W., Osborne, A., Simmons, B., & Strong, D. (2006). Construct Validity of the Balloon Analogue Risk Task (BART): Relationship with MDMA Use by Inner-City Drug Users in Residential Treatment. *Journal of Psychopathology and Behavioral Assessment*, *28*(2), 95–101.

Howard, M., & Kahana, M. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Howard, M., MacDonald, C., Tiganj, Z., Shankar, K., Du, Q., Hasselmo, M., & Eichenbaum, H. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(13), 4692–4707.

Howard, M., Shankar, K., Aue, W., & Criss, A. (2015). A distributed representation of internal time. *Psychological Review*, *122*(1), 24–53.

Hunt, M., Hopko, D., Bare, R., Lejuez, C., & Robinson, E. (2005). Construct validity of the Balloon Analog Risk Task (BART): associations with psychopathy and impulsivity. *Assessment*, *12*(4), 416–428.

Hutchison, K., Balota, D., & Ducheck, J. (2010). The utility of Stroop task switching as a marker for early-stage Alzheimer's disease. *Psychology and Aging*, *25*(3), 545–559.

Huys, Q., Maia, T., & Frank, M. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry*, *167*(7), 748–751.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263.

Kirkpatrick, R., Turner, B., & Sederberg, P. (2019). *Equal evidence perceptual tasks suggest key role for interactive competition in decision-making*. https://doi.org/10.31234/osf.io/na35q

Kopp, B., Rist, F., & Mattler, U. (1996). N200 in the flanker task as a neurobehavioral tool for investigating executive control. *Psychophysiology*, *33*(3), 282–294.

Lejuez, C., Aklin, W., Daughters, S., Zvolensky, M., Kahler, C., & Gwadz, M. (2007). Reliability and validity of the youth version of the Balloon Analogue Risk Task (BART-Y) in the assessment of risk-taking behavior among inner-city adolescents. *Journal of Clinical Child and Adolescent Psychology*, *36*(1), 106–111.

Lejuez, C., Aklin, W., Jones, H., Richards, J., Strong, D., Kahler, C., & Read, J. (2003). The Balloon Analogue Risk Task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology*, *11*(1), 26–33.

Lejuez, C., Read, J., Kahler, C., Richards, J., Ramsey, S., Stuart, G., Strong, D., & Brown, R. (2002). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology. Applied*, *8*(2), 75–84.

Light, L. L., Patterson, M. M., Chung, C., & Healy, M. R. (2004). Effects of repetition and response deadline on associative recognition in young and older adults. In *Memory & Cognition* (Vol. 32, Issue 7, pp. 1182–1193). https://doi.org/10.3758/bf03196891

Madden, D., Spaniol, J., Costello, M., Bucur, B., White, L., Cabeza, R., Davis, S., Dennis, N., Provenzale, J., & Huettel, S. (2009). Cerebral white matter integrity mediates adult age differences in cognitive performance. *Journal of Cognitive Neuroscience*, *21*(2), 289–302.

Maia, T. (2015). Introduction to the Series on Computational Psychiatry. *Clinical Psychological Science*, *3*(3), 374–377.

Mayes, A., Montaldi, D., & Migo, E. (2007). Associative memory and the medial temporal lobes. *Trends in Cognitive Sciences*, *11*(3), 126–135.

Mesulam, M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology*, *28*(5), 597–613.

Mesulam, M. (1999). Spatial attention and neglect: parietal, frontal and cingulate contributions to the mental representation and attentional targeting of salient extrapersonal events. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *354*(1387), 1325–1346.

Mulder, M., Bos, D., Weusten, J., van Belle, J., van Dijk, S., Simen, P., van Engeland, H., & Durston, S. (2010). Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological Psychiatry*, *68*(12), 1114–1119.

Müller, N., Bartelt, O., Donner, T., Villringer, A., & Brandt, S. (2003). A physiological correlate of the "Zoom Lens" of visual attention. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *23*(9), 3561–3565.

Navarro, D., & Fuss, I. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: tests of an associative deficit hypothesis. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(5), 1170–1187.

Nigbur, R., Schneider, J., Sommer, W., Dimigen, O., & Stürmer, B. (2015). Ad-hoc and context-dependent adjustments of selective attention in conflict control: an ERP study with visual probes. *NeuroImage*, *107*, 76–84.

Nunez, M., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of Mathematical Psychology*, *76*(Pt B), 117–130.

Petzschner, F., Weber, L., Gard, T., & Stephan, K. (2017). Computational Psychosomatics and Computational Psychiatry: Toward a Joint Framework for Differential Diagnosis. *Biological Psychiatry*, *82*(6), 421–430.

Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology. General*, *146*(5), 722–745.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & Starns, J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719.

Roth, M., Tym, E., Mountjoy, C., Huppert, F., Hendrie, H., Verma, S., & Goddard, R. (1986). CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *The British Journal of Psychiatry*, *149*, 698–709.

Samuel, I., Wang, C., Burke, S., Kluger, B., & Ding, M. (2019). Compensatory Neural Responses to Cognitive Fatigue in Young and Older Adults. *Frontiers in Neural Circuits*, *13*. https://doi.org/10.3389/fncir.2019.00012

Sarter, M., Berntson, G., & Cacioppo, J. (1996). Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. *The American Psychologist*, *51*(1), 13–21.

Scherbaum, S., Fischer, R., Dshemuchadse, M., & Goschke, T. (2011). The dynamics of cognitive control: evidence for within-trial conflict adaptation from frequency-tagged EEG. *Psychophysiology*, *48*(5), 591–600.

Schonberg, T., Fox, C., & Poldrack, R. (2011). Mind the gap: bridging economic and naturalistic risk-taking with cognitive neuroscience. *Trends in Cognitive Sciences*, *15*(1), 11–19.

Schubert, A., Nunez, M., Hagemann, D., & Vandekerckhove, J. (2019). Individual differences in cortical processing speed predict cognitive abilities: a model-based cognitive neuroscience account. *Computational Brain & Behavior*, *2*(2), 64–84.

Shadlen, M., & Newsome, W. (1996). Motion perception: seeing and deciding. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(2), 628–633.

Shadlen, M., & Newsome, W. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shankar, K., & Howard, M. (2012). A scale-invariant internal representation of time. *Neural Computation*, *24*(1), 134–193.

Stephan, K., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Strauss, E., Sherman, E., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. American Chemical Society.

Ter Braak, C. (2006). A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, *16*(3), 239–249.

Tiganj, Z., Cromer, J., Roy, J., Miller, E., & Howard, M. (2018). Compressed Timeline of Recent Experience in Monkey Lateral Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *30*(7), 935–950.

Tootell, R., Hadjikhani, N., Hall, E., Marrett, S., Vanduffel, W., Vaughan, J., & Dale, A. (1998). The retinotopy of visual spatial attention. *Neuron*, *21*(6), 1409–1422.

Trepel, C., Fox, C., & Poldrack, R. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Brain Research. Cognitive Brain Research*, *23*(1), 34–50.

Tsetsos, K., Gao, J., McClelland, J., & Usher, M. (2012). Using Time-Varying Evidence to Test Models of Decision Dynamics: Bounded Diffusion vs. the Leaky Competing Accumulator Model. *Frontiers in Neuroscience*, *6*, 79.

Tulving, E. (1983). Ecphoric processes in episodic memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *302*(1110), 361–371.

Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Turner, B., & Sederberg, P. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology*, *56*(5), 375–385.

Turner, B., & Sederberg, P. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250.

Turner, B., Sederberg, P., Brown, S., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384.

Tversky, A., & Kahneman, D. (2000). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Choices, Values, and Frames*, 44–66.

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Usher, M., & McClelland, J. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*(3), 757–769.

Wallsten, T., Pleskac, T., & Lejuez, C. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review*, *112*(4), 862–880.

Weichart, E., Turner, B., & Sederberg, P. (2020). A model of dynamic, within-trial conflict resolution for decision making. *Psychological Review*. https://doi.org/10.1037/rev0000191

Weichart, E. R., & Sederberg, P. B. (2020). Individual differences in attention allocation during a two-dimensional inhibitory control task. *Attention, Perception & Psychophysics*. https://doi.org/10.3758/s13414-020-02160-6

Weigard, A., Sathian, K., & Hampstead, B. (2020). Model-based assessment and neural correlates of spatial memory deficits in mild cognitive impairment. *Neuropsychologia*, *136*, 107251.

White, C., Ratcliff, R., & Starns, J. (2011). Diffusion models of the flanker task: discrete versus gradual attentional selection. *Cognitive Psychology*, *63*(4), 210–238.

White, T., Lejuez, C., & de Wit, H. (2008). Test-retest characteristics of the Balloon Analogue Risk Task (BART). *Experimental and Clinical Psychopharmacology*, *16*(6), 565–570.

Wiecki, T., Antoniades, C., Stevenson, A., Kennard, C., Borowsky, B., Owen, G., Leavitt, B., Roos, R., Durr, A., Tabrizi, S., & Frank, M. (2016). A Computational Cognitive Biomarker for Early-Stage Huntington's Disease. *PLOS ONE*, *11*(2), e0148409.

Wiecki, T., Poland, J., & Frank, M. (2015). Model-Based Cognitive Neuroscience Approaches to Computational Psychiatry. *Clinical Psychological Science*, *3*(3), 378–399.