Psychon Bull Rev (2017) 24:1234–1251
DOI 10.3758/s13423-016-1199-y

THEORETICAL REVIEW

The drift diffusion model as the choice rule in reinforcement learning

Mads Lund Pedersen[1,2] · **Michael J. Frank**[3] · **Guido Biele**[1,4]

Published online: 13 December 2016

Abstract Current reinforcement-learning models often assume simplified decision processes that do not fully reflect the dynamic complexities of choice processes. Conversely, sequential-sampling models of decision making account for both choice accuracy and response time, but assume that decisions are based on static decision values. To combine these two computational models of decision making and learning, we implemented reinforcement-learning models in which the drift diffusion model describes the choice process, thereby capturing both within- and across-trial dynamics. To exemplify the utility of this approach, we quantitatively fit data from a common reinforcement-learning paradigm using hierarchical Bayesian parameter estimation, and compared model variants to determine whether they could capture the effects of stimulant medication in adult patients with attention-deficit hyperactivity disorder (ADHD). The model with the best relative fit provided a good description of the learning process, choices, and response times. A parameter recovery experiment showed that the hierarchical Bayesian modeling approach enabled accurate estimation of the model parameters. The model approach described here, using simultaneous estimation of reinforcement-learning and drift diffusion model parameters, shows promise for revealing new insights into the cognitive and neural mechanisms of learning and decision making, as well as the alteration of such processes in clinical groups.

Keywords Decision making · Reinforcement learning · Bayesian modeling · Mathematical models

Computational models have greatly contributed to bridging the gap between behavioral and neuronal accounts of adaptive functions such as instrumental learning and decision making (Forstmann & Wagenmakers, 2015). The discovery that learning is driven by phasic bursts of dopamine coding a reward prediction error can be traced to reinforcement-learning (RL) models (Glimcher, 2011; Montague, Dayan, & Sejnowski, 1996; Rescorla & Wagner, 1972). Similarly, the current understanding of the neural mechanisms of simple decision making closely resembles the processes modeled in sequential-sampling models of decision making (Smith & Ratcliff, 2004).

RL models have been extended to account for the complexities of learning—for example, by proposing different valuations (Ahn, Busemeyer, Wagenmakers, & Stout, 2008; Busemeyer & Stout, 2002) and updating of gains and losses (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Gershman, 2015), by accounting for the role of working memory during learning (Collins & Frank, 2012), and by introducing adaptive learning rates (Krugel, Biele, Mohr, Li, & Heekeren, 2009). In contrast, the choice process during instrumental learning in RL is typically modeled with simple choice rules such as the softmax logistic function (Luce, 1959), which do not capture the dynamics of decision making (and

Electronic supplementary material The online version of this article (doi:10.3758/s13423-016-1199-y) contains supplementary material, which is available to authorized users.

✉ Mads Lund Pedersen
m.l.pedersen@psykologi.uio.no

✉ Guido Biele
guido.biele@neuro-cognition.org

[1] Department of Psychology, University of Oslo, Oslo, Norway

[2] Intervention Centre, Oslo University Hospital, Rikshospitalet, Oslo, Norway

[3] Department of Cognitive, Linguistic & Psychological Sciences, Brown Institute for Brain Science, Brown University, Providence, Rhode Island, USA

[4] Norwegian Institute of Public Health, Oslo, Norway

hence are unable to account for choice latencies). Conversely, these complexities are described well by the class of sequential-sampling models of decision making, which includes the drift diffusion model (DDM; Ratcliff, 1978), the linear ballistic accumulator model (Brown & Heathcote, 2008), the leaky competing accumulator model (Usher & McClelland, 2001), and decision field theory (Busemeyer & Townsend, 1993). The DDM of decision making is a widely used sequential-sampling model (Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff & McKoon, 2008; Wabersich & Vandekerckhove, 2013; Wiecki, Sofer, & Frank, 2013), which assumes that choices are made by continuously sampling noisy decision evidence accumulating until a decision boundary is reached in favor of one of two alternatives. The key advantage of sequential-sampling models like the DDM is that they extract more information from choice data by simultaneously fitting response time (RT; and the distributions thereof) and accuracy (or choice direction) data. Combining the dynamic learning processes across trials modeled by RL with the fine-grained account of decision processes within trials afforded by sequential-sampling models could therefore provide a richer description and new insights into decision processes in instrumental learning.

To draw on the advantages of both RL and sequential-sampling models, the goal of this article is to construct a combined model that can improve understanding of the joint latent learning and decision processes in instrumental learning. A similar approach has been described by Frank et al. (2015), who modeled instrumental learning by combining Bayesian updating as a learning mechanism with the DDM as a choice mechanism. The innovation of the research described here is that we combined a detailed description of both RL and choice processes, allowing for simultaneous estimation of their parameters. The benefit of using the DDM as the choice rule in an RL model is that a combined model can capture various factors, including the sensitivity to expected rewards, how they are updated by prediction errors, and the trade-off between speed versus accuracy during response selection. This endeavor can help decompose mechanisms of choice and learning in a richer way than could be accomplished by either RL or DDM models alone, while also laying the groundwork to further investigate the neural underpinnings of these subprocesses by fitting model parameters based on neural regressors (Cavanagh, Wiecki, & Cohen, 2011; Frank et al., 2015).

One hurdle for the implementation of complex models of learning and decision making has traditionally been the difficulty to fit models with a large number of parameters. The advancement of methods for Bayesian parameter estimation in hierarchical models has helped address this problem (Lee & Wagenmakers, 2014; Wiecki et al., 2013). A hierarchical Bayesian approach improves the estimation of individual parameters by assuming that the parameters for individuals are drawn from group distributions (Kruschke, 2010), yielding mutually constrained estimates of group and individual parameters that can improve parameter recovery for individual subjects (Gelman et al., 2013).

In the following sections, we will describe RL models and the DDM in detail, before explaining and justifying a combined model. We will propose potential mechanisms involved in instrumental learning, and describe models expressing these mechanisms. Next, we compare how well these models describe data from an instrumental-learning task in humans. To show that combining RL and the DDM is able to account for data and provide new insight, we will demonstrate that the best-fitting model can disentangle effects of stimulant medication on learning and decision processes in attention-deficit hyperactivity disorder (ADHD). Finally, to ensure that the model parameters capture the submechanisms they are intended to describe, we show that the generated parameters can successfully be recovered from simulated data.

Reinforcement-learning models

RL models were developed to describe associative and instrumental learning (Bush & Mosteller, 1951; Rescorla & Wagner, 1972). The central tenet to these models is that learning is driven by unexpected outcomes—for example, the surprising occurrence or omission of reward, in associative learning, or when an action results in a larger or smaller reward than is expected, in instrumental learning. An unexpected event is captured by the prediction error (PE) signal, which describes the difference between the observed and predicted rewards. The PE signal thus generates an updated reward expectation by correcting past expectations.

The strong interest in RL in cognitive neuroscience was amplified by the finding that the reward PE is signaled by midbrain dopaminergic neurons (Montague et al., 1996; Schultz, Dayan, & Montague, 1997), which can then alter expectations and subsequent choices by modifying synaptic plasticity in the striatum (see Collins & Frank, 2014, for a review and models).

RL models typically consist of, at least, an updating mechanism for adapting reward expectations of choice options, and an action selection policy that describes how choices between options are made. A popular learning algorithm is the delta learning rule (Bush & Mosteller, 1951; Rescorla & Wagner, 1972), which can be used to describe trial-by-trial instrumental learning. According to this algorithm, the reward value expectation for the chosen option i

on trial t, $V_i(t)$, is calculated by summing the reward expectation from the previous trial and the reward PE:

$$V_i(t) = V_i(t-1) + \eta[\text{Reward}_i(t-1) - V_i(t-1)]. \quad (1)$$

The PE is weighted with a learning rate parameter η, such that larger learning rates close to 1 lead to fast adaptation of reward expectations, and small learning rates near 0 lead to slow adaptation. The process of choosing between options can be described by the softmax choice rule (Luce, 1959). This choice rule models the probability $p_i(t)$ that a decision maker will choose one option i among all options j:

$$p_i(t) = \frac{e^{(\beta(t) \times V_i(t))}}{\sum_{j=1}^{n} e^{[\beta(t) \times V_j(t)]}}. \quad (2)$$

The parameter β governs the sensitivity to rewards and the exploration–exploitation trade-off. Larger values indicate greater sensitivity to the rewards received, and hence more deterministic choice of options with higher reward values. Following Busemeyer and colleagues' assumptions in the expectancy valence (EV) model (Busemeyer & Stout, 2002), sensitivity can change over the course of learning following a power function:

$$\beta(t) = (t/10)^c, \quad (3)$$

where consistency c is a free parameter describing the change in sensitivity. Sensitivity to expected rewards increases during the course of learning when c is positive, and decreases when c is negative. Change in sensitivity is related to the exploration–exploitation trade-off (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Sutton & Barto, 1998), in which choices are at first typically driven more by random exploration, but then gradually shift to exploitation in stable environments when decision makers learn the expected values and preferentially choose the option with the highest expected reward. The EV model (Busemeyer & Stout, 2002) thus assumes that the consistency of choices with learned values is determined by one decision variable. Reward sensitivity normally increases (and exploration decreases) with learning, but it can also remain stable or even decrease, due to boredom or fatigue. Other RL models, such as the prospect valence learning model, assume a trial-independent reward sensitivity in which reward sensitivity remains constant throughout learning (Ahn et al., 2008; Ahn, Krawitz, Kim, Busemeyer, & Brown, 2011).

We hypothesized that the β sensitivity decision variable captures potentially independent decision processes that can be disentangled using sequential sampling models, such as the DDM, which incorporate the full RT distribution of choices. For example, frequent choosing of superior options can result from clear and accurate representations of the option values, from favoring accurate over speedy choosing, or from a tendency to favor exploitation over exploration. Conversely, frequent choosing of inferior options can be caused by noisy and biased representations of the option values, by a focus on speedy over accurate choosing, or by the exploration of alternative options with lower but uncertain payoff expectations. Using the DDM as the choice function in an RL model can help to disentangle the representations of option values from a focus on speed versus accuracy (due to their differential influences on the RT distributions), and thus improve knowledge of the latent processes involved in choosing during reinforcement-based decision making.

Drift diffusion model of decision making

The DDM is one instantiation of the broader class of sequential-sampling models used to quantify the processes underlying two-alternative forced choice decisions (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Jones & Dzhafarov, 2014; Ratcliff, 1978; Ratcliff & McKoon, 2008; Smith & Ratcliff, 2004). The DDM assumes that decisions are made by continuously sampling noisy decision evidence until a decision boundary in favor of one of two alternatives is reached (Ratcliff & Rouder, 1998). Consider deciding whether a subset of otherwise randomly moving dots are moving left or right, as in a random dot-motion task (Shadlen & Newsome, 2001). Such a decision process is represented in Fig. 1 by sample paths with a starting point indicated by the parameter z. The difference in evidence between dot-motion toward the left or toward the right is continuously gathered until a boundary for one of the two alternatives (upper or lower boundary, here representing "left" and "right") is reached. According to the DDM, accuracy and RT distributions depend on a number

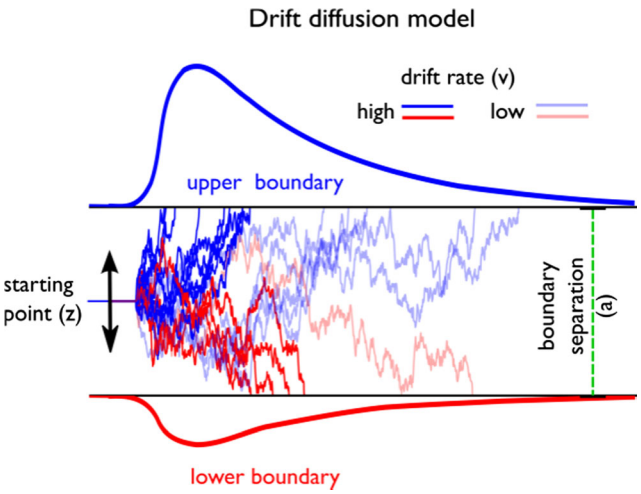

Fig. 1 Main features of the drift diffusion model. The accumulation of evidence begins at a starting point (z). Evidence is represented by sample paths with added Gaussian noise, and is gathered until a decision boundary is reached (upper or lower) and a response is initiated. High and low drift rates are depicted as lines with high and low color saturation, respectively. From "HDDM: Hierarchical Bayesian Estimation of the Drift-Diffusion Model in Python," by Wiecki, Sofer, and Frank, 2013, *Frontiers in Neuroinformatics*, *7*, Article 14. Copyright 2013 by Frontiers Media S.A. Adapted with permission

of decision parameters. The drift rate (v) reflects the average speed with which the decision process approaches the response boundaries. High drift rates lead to faster and more accurate decisions. The boundary separation parameter (a), which adjusts the speed–accuracy trade-off, describes the amount of evidence needed until a decision threshold is reached. Wider decision boundaries lead to slower and more accurate decisions, whereas narrower boundaries lead to faster but more error-prone decisions. The starting point (z) represents the extent to which one decision alternative is preferred over the other, before decision evidence is available—for example, because of a higher occurrence or incentive value of this alternative. The nondecision time parameter (T_{er}) captures the time taken by stimulus encoding and motor processes. The full DDM also includes parameters that capture between-trial variability in the starting point, drift rate, and nondecision time (Ratcliff & McKoon, 2008). To keep the model reasonably simple, these will not be included in our combined model. Although the DDM was initially developed to describe simple perceptual or recognition decisions (Ratcliff, 1978; Ratcliff & Rouder, 1998), there is a strong research tradition of using sequential-sampling models to explain value-based choices, including multiattribute choice and risky decision making (Busemeyer & Townsend, 1993; Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2001). One of the earliest applications of sequential-sampling models for value-based choice (Busemeyer, 1985) also investigated decision making in a learning context. Unlike the present research, this research did not explicitly model the learning process, but focused instead on the qualitative predictions of sequential-sampling models as opposed to fitting model parameters.

Importantly, the DDM and, more broadly, sequential-sampling models of decision making not only successfully describe perceptual and value-based decisions in both healthy and clinical populations (White, Ratcliff, Vasey, & McKoon, 2010), but are also consistent with the neurobiological mechanisms of decision making uncovered in neurophysiological and neuroimaging experiments (Basten, Biele, Heekeren, & Fiebach, 2010; Cavanagh et al., 2011; Cavanagh, Wiecki, Kochar, & Frank, 2014; Forstmann et al., 2011; Frank et al., 2015; Hare, Schultz, Camerer, O'Doherty, & Rangel, 2011; Kayser, Buchsbaum, Erickson, & D'Esposito, 2010; Krajbich & Rangel, 2011; Mulder, van Maanen, & Forstmann, 2014; Nunez, Srinivasan, & Vandekerckhove, 2015; Pedersen, Endestad, & Biele, 2015; Ratcliff, Cherian, & Segraves, 2003; Smith & Ratcliff, 2004; Turner, van Maanen, & Forstmann, 2015; Usher & McClelland, 2001).

Reinforcement learning drift diffusion model

The rationale for creating a reinforcement learning drift diffusion (RLDD) model is to exploit the DDMs ability to account for the complexities of choice processes during instrumental learning. A model describing the results of an instrumental learning task in a DDM framework could be expressed in several ways. We will therefore start with a basic model of the decision and learning processes in instrumental learning, and then describe alternative expressions of these processes. Furthermore, we will compare how well the models fit data from a common probabilistic RL task, and examine through a posterior predictive check how well the best-fitting model describes the observed choices and RT distributions.

The DDM calculates the likelihood of the RT of a choice x with the Wiener first-passage time (WFPT) distribution,

$$RT(x) \sim \mathrm{WFPT}[a, T_{er}, z, v(t)], \tag{4}$$

where the WFPT returns the probability that x is chosen with the observed RT. In this basic RLDD model, the nondecision time T_{er}, starting point z, and boundary separation a are trial-independent free parameters, as in the ordinary DDM. The drift rate $v(t)$ varies from trial to trial as a function of the difference in the expected rewards, multiplied by a scaling parameter m, which can capture differences in the ability to use knowledge of the reward probabilities:

$$v(t) = \left[V_{\mathrm{upper}}(t) - V_{\mathrm{lower}}(t)\right] \times m. \tag{5}$$

$V_{\mathrm{upper}}(t)$ and $V_{\mathrm{lower}}(t)$ represent the reward expectations for the two response options. The scaling parameter also ensures that $V - \Delta = V_{\mathrm{upper}} - V_{\mathrm{lower}}$ is transformed to an appropriate scale in the DDM framework. V values are initialized to 0 and updated as a function of the reward PEs at a rate dependent on a free parameter, learning rate η (Eq. 1). In this basic model, choice sensitivity is constant over time, meaning that equal differences in V values will lead to the same drift rates, independent of reward history (i.e., the exploitation–exploration trade-off does not change over the course of learning). This approach has previously been applied successfully to instrumental-learning data in a DDM framework, in a model that used Bayesian updating as a learning mechanism with no additional free parameters (Frank et al., 2015). In contrast to RL models, Bayesian updating implies a reduced impact of feedback later during learning, because the distribution of prior expectations becomes narrower as more information is incorporated. However, such a model does not allow one to estimate the variants of RL that might better describe human learning.

Drift rate

Whereas the basic model assumes constant sensitivity to payoff differences, it has been shown that sensitivity can increase or decrease over the course of learning—for example, due to increased fatigue or certainty in reward expectations, or due to changes in the tendency to exploit versus explore (Busemeyer

& Stout, 2002). Accordingly, an extended drift rate calculation allows for additional variability by assuming that the multiplication factor of V-Δ changes according to a power function:

$$v(t) = \left(V_{\text{upper}} - V_{\text{lower}}\right) \times (t/10)^{p}. \quad (6)$$

In this expression, the drift rate increases throughout learning when the choice consistency parameter p is positive, representing increased confidence in the learned values, and decreases when p is negative, representing boredom or fatigue. The power function could also account for a move from exploration to exploitation when reward sensitivity increases.

Boundary separation

In the basic model described above, the boundary separation (sometimes referred to as the *decision threshold*) is assumed to be static. However, it could be that the boundary separation is altered as learning progresses. Time-dependent changes of decision thresholds could follow a power function by calculating the threshold as a combination of a boundary baseline bb times the boundary power parameter bp multiplied by trial t:

$$a(t) = bb \times (t/10)^{bp}. \quad (7)$$

Learning rate

Several studies have reported differences in updating of expected rewards following positive and negative PEs (Gershman, 2015), which is hypothesized to be caused by the differential roles of striatal D1 and D2 dopamine receptors in separate corticostriatal pathways (Collins & Frank, 2014; Cox, Frank, Larcher, Fellows, & Clark, 2015; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007). We therefore assumed that V values could be modeled with asymmetric updating rates, where η^{+} and η^{-} are used to update the expected rewards following positive and negative PEs, respectively.

Model selection

The various mechanisms for drift rate, boundary separation, and learning rate outlined above can be combined into different models, and these models can be compared for their abilities to describe data. Identifying a model with a good fit requires several considerations (Heathcote, Brown, & Wagenmakers, 2015). First of all, a model describing latent cognitive processes needs to be able to fit data from human or animal experiments. To separate learning from choice sensitivities, we therefore fit models on data from subjects performing a probabilistic instrumental-learning task. To determine which model described the data best, we first compared models on their relative fits to the data, and then further ascertained the validity of the best-fitting model by examining its absolute fit (Steingroever, Wetzels, & Wagenmakers, 2014) through posterior predictive checks (Gelman, Meng, & Stern, 1996). A useful model should also have clearly interpretable parameters, which in turn depends on the ability to recover the model parameters—that is, the model must be possible to identify the generative parameters. We therefore performed a parameter recovery experiment as a final step to ensure that the fitted parameters described the processes that we propose they describe.

Method

Instrumental-learning task

The probabilistic selection task (PST) is an instrumental-learning task that has been used to describe the effect of dopamine on learning in both clinical and normal populations (Frank, Santamaria, O'Reilly, & Willcutt, 2007; Frank, Seeberger, & O'Reilly, 2004), in which increases in dopamine boost relative learning from positive as compared to negative feedback. On the basis of a detailed neural-network model of the basal ganglia, these effects are thought to be due to the selective modulation of striatal D1 and D2 receptors through dopamine (Frank et al., 2004). The task has been used to investigate the effects of dopamine on learning and decision making in ADHD (Frank, Santamaria, et al., 2007), autism spectrum disorder (Solomon, Frank, & Ragland, 2015), Parkinson's disease (Frank et al., 2004), and schizophrenia (Doll et al., 2014), among others.

The PST consists of a learning phase and a test phase. During the learning phase, decision makers are presented with three different stimulus pairs (AB, CD, EF), represented as Japanese hiragana letters, and learn to choose one of the two stimuli in each pair on the basis of reward feedback. Reward probabilities differ between the stimulus pairs. In AB trials, choosing A is rewarded with a probability of .8, whereas B is rewarded with a probability of .2. In the CD pair, C is rewarded with a probability of .7, and D .3, and in the EF pair, E is rewarded with a probability of .6, and F .4. Because stimulus pairs are presented in random order, the reward probabilities for all six stimuli have to be maintained throughout the task. Success in the learning phase is to learn to maximize rewards by choosing the optimal (A, C, E) over the suboptimal (B, D, F) option in each stimulus pair (AB, CD, EF). Subjects perform as many blocks (of 60 trials each) as required until their running accuracy at the end of a block is above 65% for AB pairs, 60% for CD pairs, and 50% for EF pairs, or until they complete six blocks (360 trials) if the criteria are not met. The PST also includes a test phase, which we will not examine

in the present research because it does not involve trial-to-trial learning and exploration. Instead, we will focus on the learning phase of the PST, which can be described as a probabilistic instrumental-learning task.

The data from the learning phase of the PST in Frank, Santamaria, O'Reilly, and Willcutt (2007) were used to assess the RLDD models' abilities to account for data from human subjects. We also used the task to simulate data from synthetic subjects in order to test the best-fitting model's ability to recover the parameters. In the original article, the effects of stimulant medication were tested in ADHD patients with a within-subjects medication manipulation, and 17 ADHD subjects were also compared to 21 healthy controls. In the present study, we focused on the results from ADHD patients to understand the causes of the appreciable effects of medication on this group. Subjects were tested twice in a within-subjects design. The order of medication administration was randomized between the ADHD subjects. The results showed that medication improved learning performance, and the subsequent test phase showed that this change was accompanied by a selective boost in reward learning rather than in learning from negative outcomes, consistent with the predictions of the basal ganglia model related to dopaminergic signaling in striatum (Frank, Santamaria, et al., 2007).

Analysis

Parameters in the RLDD models were estimated in a hierarchical Bayesian framework, in which prior distributions of the model parameters were updated on the basis of the likelihood of the data given the model, to yield posterior distributions. The use of Bayesian analysis, and specifically hierarchical Bayesian analysis, has increased in popularity (Craigmile, Peruggia, & Van Zandt, 2010; Lee & Wagenmakers, 2014; Peruggia, Van Zandt, & Chen, 2002; Vandekerckhove, Tuerlinckx, & Lee, 2011; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010), due to its several benefits relative to traditional analysis. First, posterior distributions directly convey the uncertainty associated with parameter estimates (Gelman et al., 2013; Kruschke, 2010). Second, in a hierarchical approach, individual and group parameters are estimated simultaneously, which ensures mutually constrained and reliable estimates of both the group and the individual parameters (Gelman et al., 2013; Kruschke, 2010). These benefits make a Bayesian hierarchical framework especially valuable when estimating individual parameters for complex models based on a limited amount of data, as is often the case in research with clinical groups (Ahn et al., 2011) or in experiments combining parameter estimates with neural data to identify neural instantiations of proposed processes in cognitive models (Cavanagh et al., 2011). In the context of modeling decision making, Wiecki et al. (2013) showed that a Bayesian hierarchical approach recovers the parameters of the DDM better than do other methods of analysis.

We used the JAGS Wiener module (Wabersich & Vandekerckhove, 2013) in JAGS (Plummer, 2004), via the rjags package (Plummer & Stukalov, 2013) in R (R Development Core Team, 2013), to estimate posterior distributions. Individual parameters were drawn from the corresponding group-level distributions of the baseline (OFF) and medication effect parameters. Group-level parameters were drawn from uniformly distributed priors and were estimated with noninformative mean and standard deviation group priors. For each trial, the likelihood of the RT was assessed by providing the WFPT distribution with the boundary separation, starting point, nondecision time, and drift rate parameters. Responses in the PST data were accuracy-coded, and symbol–value associations were randomized across subjects. It was therefore assumed that the subjects would not develop a bias, represented as a change in starting point (z) toward a decision alternative. To examine whether learning results in a change of the starting point in the direction of the optimal response, we compared the RTs for correct and error responses in the last third of the experiment. Changes in starting point should be reflected in slower error RTs (Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012; Ratcliff & McKoon, 2008). We focused this analysis on the last third of the trials, because in those trials subjects would be more likely to maximize rewards and less likely to make exploratory choices, which more frequently are "erroneous," but could also be slower for reasons other than bias. Comparison of the median correct and error RTs showed no clear RT differences, such that the alternative hypothesis was only 1.78 times more likely than the null-hypothesis (median error RT = 1.039 [0.406] s, median correct RT = 0.935 [0.373] s, $BF_{10} = 1.78$; Morey & Rouder, 2015). Hierarchical modeling of median RTs that explicitly accounted for RT differences between the conditions showed the same results, whereas an analysis of all trials showed even weaker evidence for slower error responses ($BF_{10} = 1.37$). The starting point was therefore fixed at .5. Nonresponses (0.011%) and RTs faster than 0.2 s (1.5%) were removed prior to analysis.

To capture individual within-subjects effects of medication, we used a dummy variable coding for the medication condition, and estimated for each trial the individual parameters for OFF as a baseline and the individual parameters for ON as baseline, plus the effect of medication (see The supplementary materials for the model code). To assess the effect of medication, we report the posterior means, 95% highest density intervals, and Bayes factors as measures of evidence for the existence of directional effects. Because all priors for group effects are symmetric, Bayes factors for directional effects can simply be calculated as the ratio of the posterior mass above zero to the posterior mass below zero (Marsman & Wagenmakers, 2016).

Relative model fit

Comparison of the relative model fits was performed with an approximation to the leave-one-out (LOO) cross-validation (Vehtari, Gelman, & Gabry, 2016). In our application, LOO repeatedly excludes the data from one subject and uses the remaining subjects to predict the data for the left-out subject (i.e., subjects and not trials are independent observations). It therefore balances between the likelihood of the data and the complexity of the model, because both too-simple and too-complex models would give bad predictions, due to under- and overfitting, respectively. Higher LOO values indicate better fits. To directly compare the model fits, we computed the difference in predictive accuracy between models and its standard error, and assumed that the model with the highest predictive ability had a better fit if the 95% confidence interval of the difference did not overlap with 0.

Absolute model fit

One should not be encouraged by a relative model comparison alone (Nassar & Frank, 2016); the best-fitting model of those devised might still not capture the key data. The best-fitting model from the relative model comparison was therefore evaluated further to determine its ability to account for key features of the data by using measures of absolute model fit, which involves comparing observed results with data generated from the estimated parameters. We used two absolute model fit methods, based on the post-hoc absolute-fit method and the simulation method described by Steingroever and colleagues (Steingroever et al., 2014). The post-hoc absolute-fit method (also called *one-step-ahead prediction*) tests a model's ability to fit the observed choice patterns given the history of previous choices, whereas the simulation method tests a model's ability to generate the observed choice patterns. We used both methods because models that accurately fit observed choices do not necessarily accurately generate them (Steingroever et al., 2014). Therefore, a model with a good match to the observed choice patterns in both methods could, with a higher level of confidence, be classified as a good model for describing the underlying process. The methods were used as posterior predictive checks (Gelman et al., 2013) to identify the model's ability to re-create both the evolution of choice patterns and the full RT distribution of choices.

In the post-hoc absolute-fit method, parameter value combinations were sampled from the individuals' joint posterior. For each trial, observed payoffs were used together with learning parameters to update the expected rewards for the next trial. Expected rewards were then used together with decision parameters to generate choices for the next trial with the rwiener function from the RWiener package (Wabersich & Vandekerckhove, 2014). This procedure was performed 100 times to account for posterior uncertainty, each time drawing a parameter combination from a random position in the individuals' joint posterior distribution.

The simulation method followed the same procedure as the post-hoc absolute-fit method, with the exception that the expected values were updated with payoffs from the simulated, not from the observed, choices. The payoff scheme was the same as in the PST task, and each synthetic subject made 212 choices, which was the average number of trials completed by the subjects in the PST dataset. We accounted for posterior uncertainty in the simulation method following the same procedure as for the post-hoc absolute-fit method.

Results

Model fit of data from human subjects

Relative model fit

We compared eight models with different expressions of latent processes assumed to be involved in reinforcement-based decision making, based on data from 17 adult ADHD patients in the learning phase of the PST task (Frank, Santamaria, et al., 2007). All models were run with four chains with 40,000 burn-in samples and 2,000 posterior samples each. To assess convergence between the Markov chain Monte Carlo (MCMC) chains, we used the $\hat{R}$ statistic (Gelman et al., 1996), which measures the degree of variation between chains relative to the variation within chains. The maximum $\hat{R}$ value across all parameters in all eight models was 1.118 (four parameters in one model above 1.1), indicating that for each model all chains converged successfully (Gelman & Rubin, 1992). The LOO was computed for all models as a measure of the relative model fit.

The eight models tested differed in two expressions for calculating drift rate variability, two expressions for calculating boundary separation, and either one or two learning rates (see Table 1 for descriptions and LOO values for all tested models). All models tested had a better fit than a pure DDM model assuming no learning processes—that is, with static decision parameters (Table 1, Supplementary Table 1). The drift rate was calculated as the difference between expected rewards multiplied by a scaling parameter, which was either constant (m) or varied according to a power function (p; see above). The trial-independent scaling models (mean LOO: –16.75) provided on average a better fit than the trial-dependent scaling models (mean LOO: –16.805), although this difference was weak (confidence interval of the difference: –0.019, 0.128).

Similarly, boundary separation was modeled as a fixed parameter or varied following a power function, changing across trials. The models in which the boundary was trial-dependent had on average a better fit (mean LOO: –16.573) than the

Table 1 Summary of reinforcement-learning drift diffusion models

Model	Drift Rate Scaling	Boundary Separation	Learning Rate	elpd	Rank
1	constant	fixed	single	–17.023	7
2	constant	power	single	–16.600	3
3	power	fixed	single	–17.108	8
4	power	power	single	–16.697	4
5	constant	fixed	dual	–16.891	5
6*	constant	power	dual	–16.488	1*
7	power	fixed	dual	–16.909	6
8	power	power	dual	–16.506	2
DDM	NA	fixed	NA	–17.860	9

In the Drift Rate Scaling column, "constant" indicates that V-delta was multiplied by a constant parameter *m*, whereas "power" indicates that V-delta was multiplied with a parameter *p* following a power function. For Boundary Separation, "fixed" indicates that boundary separation was a trial-independent free parameter, whereas "power" means that boundary separation was estimated with a power function. Under Learning Rate, "dual" represents models with separate learning rates for both positive and negative prediction errors, whereas "single" models estimate one learning rate, ignoring the valence of the prediction error. The best-fitting model is marked with an asterisk. elpd = expected log pointwise predicitive density, NA = not applicable

models with fixed boundary separations (mean LOO: –16.983), and this effect was strong (confidence interval of difference: 0.172, 0.648). Separating the learning rates for positive and negative PEs (mean LOO: –16.699) resulted in an overall better fit (confidence interval of difference: 0.052, 0.264) than using a single learning rate (mean LOO: –16.857).

Pair-wise comparison of the model fits revealed that two models (Models 6 and 8) had a better fit than the other models (see Supplementary Table 1). The mean predictive accuracy of Model 6 was highest, but it could not be confidently distinguished from the fit of Model 8. Nevertheless, Model 6 had a slightly better fit (–16.488 vs. –16.506) and had all the properties favored when comparing across models, in that its drift rate was multiplied by a constant scaling factor, the boundary separation was estimated to change following a power function, and learning rates were split by the sign of the PE (Table 1). On the basis of these results, we selected Model 6 to further investigate how an RLDD model can describe data from the learning phase of the PST, and to test its ability to recover parameters from simulated data.

Absolute model fit

The four chains of Model 6 (Table 1) converged for all group- and individual-level parameters, with $\hat{R}$ values between 1.00 and 1.056 (see https://github.com/gbiele/RLDDM for model code and data). There were some dependencies between the group parameters (Fig. 2), in particular a negative correlation between drift rate scaling and the positive learning rate.

Comparing models using estimates of relative model fit such as the LOO does not assess whether the models tested are good models of the data. Absolute model fit procedures, however, can inform as to whether a model accounts for the observed results. A popular approach to estimate absolute model fit is to use posterior predictive checks (Gelman et al., 2013), which involve comparing the observed data with data generated on the basis of posterior parameter distributions. Tests of absolute model fit for RL models often include a comparison of the evolution of choice proportions (learning curves) for observed and predicted data. To validate sequential-sampling models like the DDM, however, one usually compares observed and predicted RT distributions. Because the RLLD models use both choices and RTs to estimate parameters, we created data with both the post-hoc absolute-fit method and the simulation method procedures described above (Steingroever et al., 2014), and visually compared the simulated data with the observed experimental data on measures of both RT and choice proportion (visual posterior predictive check; Gelman & Hill, 2007).

Posterior predictive plots for the development of choice proportions over time are shown in Fig. 3. For each subject, trials were grouped into bins of ten for each difficulty level. The average choice proportions in each subject's bins were then averaged to give the group averages shown in Fig. 3. The leftmost plots display the mean development of observed choice proportions in favor of the good option from each difficulty level for the OFF and ON medication conditions. The middle and rightmost plots display the mean probabilities of choosing the good option on the basis of the post-hoc absolute-fit method and the simulation method, respectively. The degree of model fit is indicated by the degree to which the generated choices resemble the observed choices. From visually inspecting the graphs, it is clear that while ON medication, the subjects on average learned to choose the correct option in all three stimulus pairs, whereas while OFF medication they did not achieve a higher accuracy than about 60% for any of the stimulus pairs. The fitted model recreates this overall pattern: Both methods identify improved performance in the ON condition, which is stronger for the more deterministic reward conditions, while also recreating the lack of learning in the OFF group. The model does not recreate the short-term fluctuations in choice proportions, which could reflect other contributions to trial-to-trial adjustments in this task, outside of instrumental learning—for example, from working memory processes (Collins & Frank, 2012). Finally, the simulation method slightly overestimated the performances in both groups.

The models described here were designed to account for underlying processes by incorporating RTs in addition to choices. Therefore, a good model should also be able to

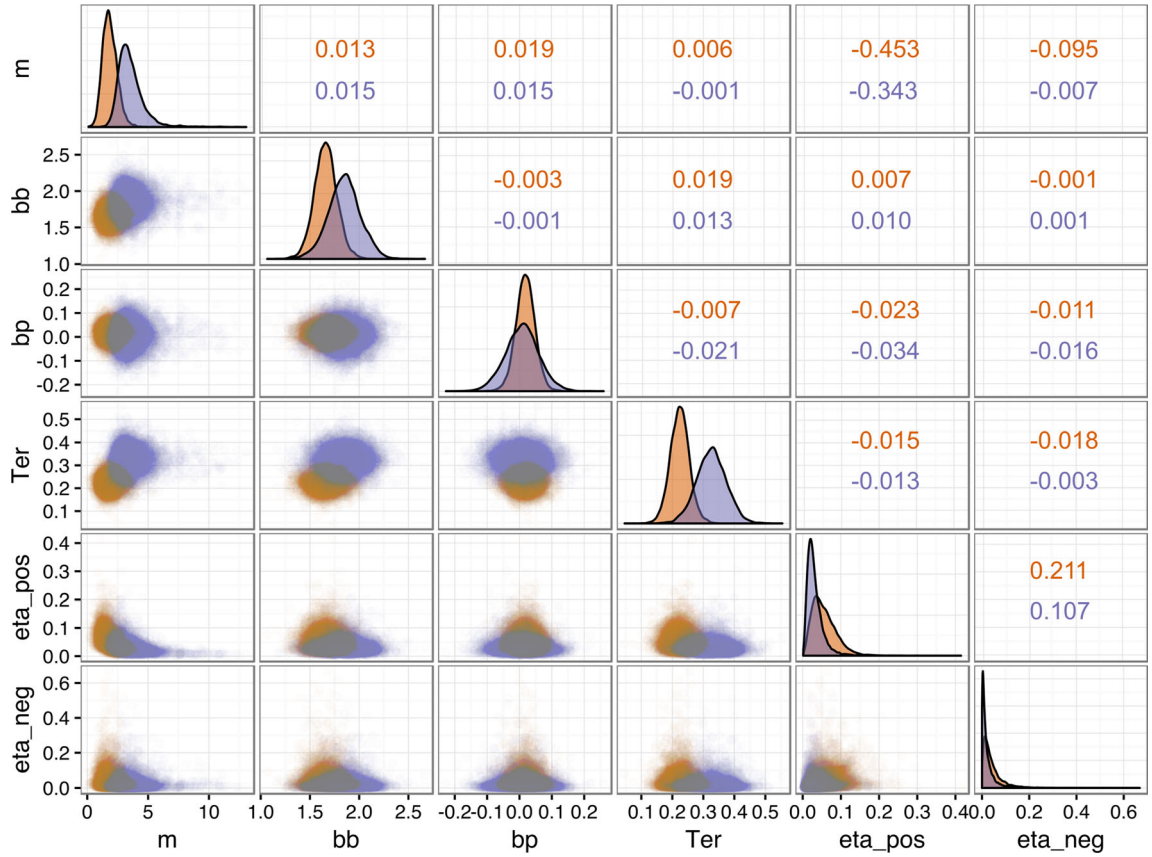

Fig. 2 Scatterplot and density of group parameter estimates from posterior distributions off (red) and on (purple) medication. bb = boundary baseline, bp = boundary power, eta_pos = learning rate for positive prediction errors (PEs), eta_neg = learning rate for negative PEs, m = drift rate scaling, T_{er} = nondecision time

predict the RT distributions of choices. The posterior predictive RTs of choices based on the post-hoc absolute-fit method and the simulation method are shown in Fig. 4 as densities superimposed on histograms of the observed results. Responses in favor of suboptimal options are coded as negative RTs. The results from both the post-hoc absolute-fit method and the simulation method re-create the result that overall accuracy is higher in the easier conditions ON medication, while slightly overestimating the proportion of correct trials OFF medication. The tails of the distributions are accurately predicted for all difficulty levels for both groups.

Effects of medication on learning and decision mechanisms

We estimated group and individual parameters dependent on the medication manipulation, both to test whether the results from the model are in line with the behavioral results reported by Frank, Santamaria, O'Reilly, and Willcutt (2007) and to examine whether they can contribute to a mechanistic explanation of the processes driving the effects of stimulant medication in ADHD (see the Discussion section). Group-level parameters of the within-subjects medication effects were used to assess how the stimulant medication influenced performance (Fig. 5 and Table 2; Wetzels & Wagenmakers, 2012).

Following Jeffreys's evidence categories for Bayes factors (Jeffreys, 1998), the within-subjects comparison revealed strong or very strong evidence that medication increased the drift rate scaling, nondecision time, and boundary separation (Fig. 5). The results also indicated substantial evidence that medication led to lower positive and negative learning rates.

Parameter recovery from simulated data

As a validation of the best-fitting RLDD model (Model 6, Table 1), we performed a parameter recovery study by estimating the posterior distributions of the parameters on simulated data. We used estimated parameter values from the original PST data to select plausible values for the free parameters in the best-fitting model. Assigning three values for each of the six parameters resulted in a matrix with $3^6 = 729$ unique combinations of parameter values. The choice and RT data were simulated using

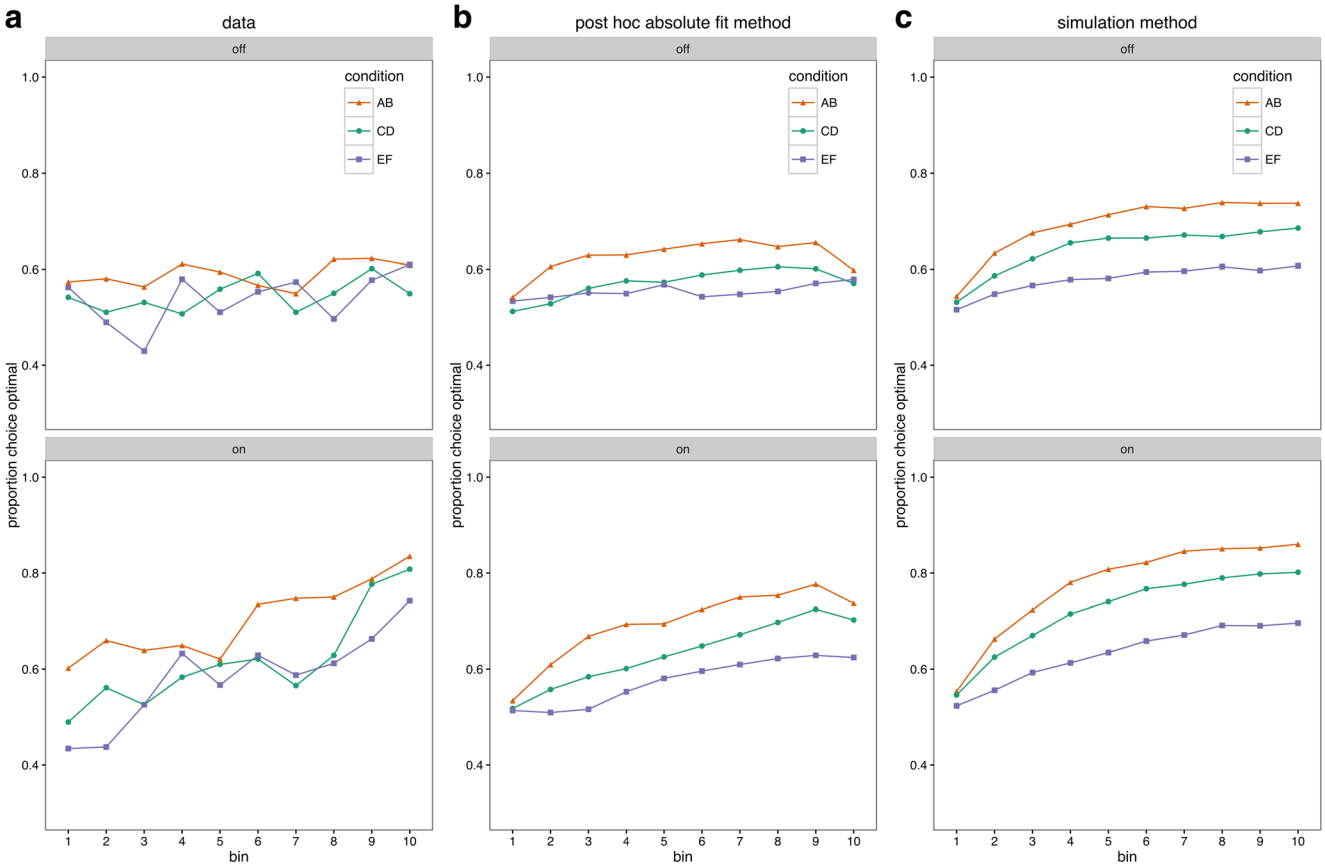

Fig. 3 Development of mean proportions of choices in favor of the optimal option for the OFF (top row) and ON (bottom row) medication groups, for (**a**) the observed data, (**b**) the post-hoc absolute-fit method, and (**c**) the simulation fit method, across stimulus pairs AB, CD, and EF (see panel legends), which had reward probabilities for the optimal and suboptimal options of .8–.2, .7–.3, and .6–.4, respectively. The choices were fit (**b**) and simulated (**c**) by drawing 100 samples from each subject's posterior distribution. For each subject, trials were grouped into bins of ten for each difficulty level, and group averages were created from the individual mean choice proportions for each bin

assigned parameter values. For each of the 729 combinations of parameter values, we created data for five synthetic subjects performing 212 trials each (the mean number of trials completed in the PST dataset).

The model was run with two MCMC chains with 2,000 burn-in samples and 2,000 posterior samples for each chain. The $\hat{R}$ values for the posterior distributions indicated convergence, with point estimate values of $\hat{R}$ between 1 and 1.15 for

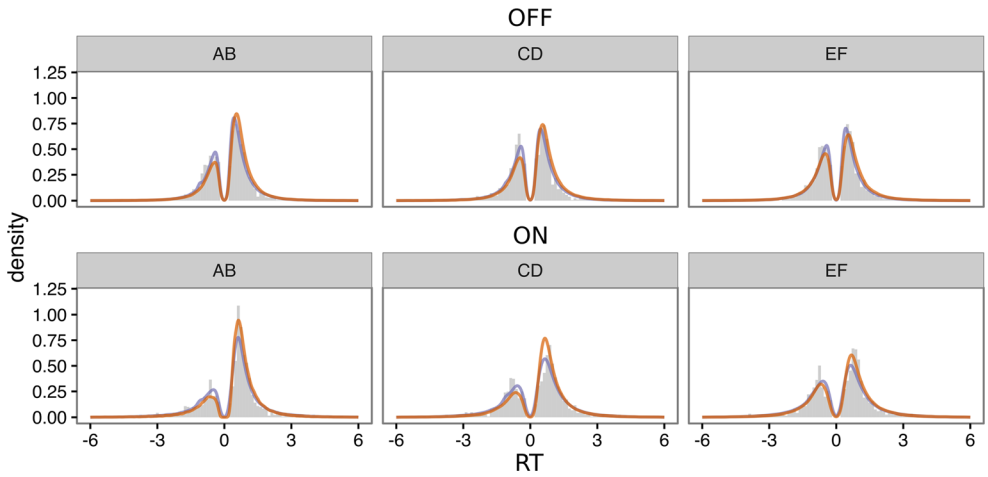

Fig. 4 Posterior predictive RT distributions across stimulus pairs, shown separately for the OFF and ON medication groups. Gray histograms display the observed results, and density lines represent generated results from the post-hoc absolute-fit method and the simulation fit method (in red and purple online, respectively). Choices in favor of the suboptimal option are coded as negative

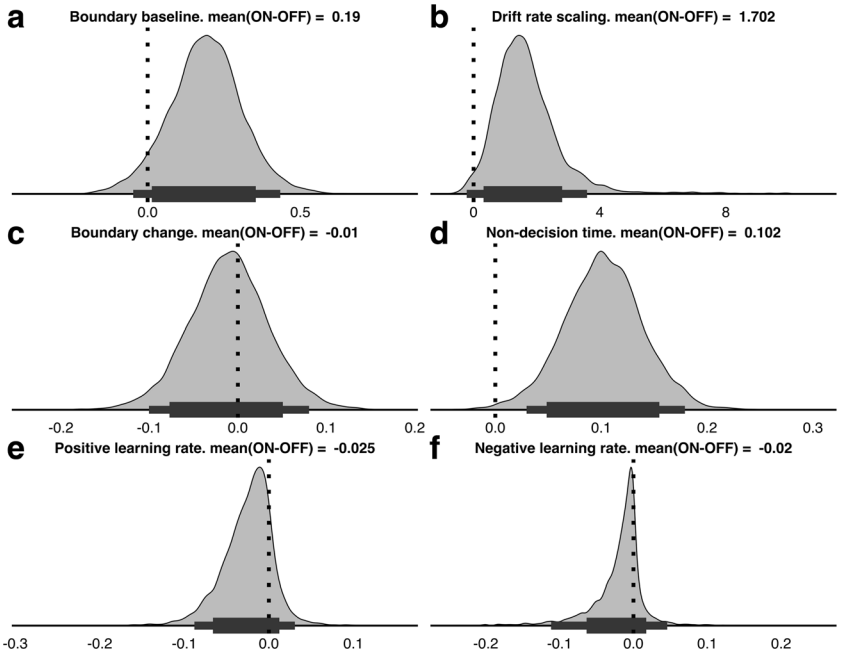

Fig. 5 Posterior distributions of differences for ADHD subjects on versus off stimulant medication, for boundary separation (**a**), drift rate scaling (**b**), boundary change (**c**), nondecision time (**d**), and positive (**e**) and negative (**f**) learning rates. Thick and thin horizontal bars below the distributions represent the 85% and 95% highest density intervals, respectively

all group and individual parameter estimations, with the exception of two parameter estimates of $\hat{R}$ at 1.20 and 1.235. Figure 6 displays the means of the posterior distributions for the estimated parameters, together with the simulated parameter values. The figure shows that the model successfully recovered the parameter values from the simulations, with means of the posterior distribution that were close to the simulated values across all parameters. The exception was the estimated learning rates, in which the highest and lowest generated values were somewhat under- and overestimated, respectively, but were still estimated to be higher and lower than the estimations for the other generated values (i.e., there was a strong correlation between the generated and estimated parameters). There were generally weak dependencies between the parameters, with low correlations between the mean parameter estimates (Fig. 7).

Discussion

We proposed a new integration of two popular computational models of reinforcement learning and decision making. The key innovation of our research is to implement the DDM as the choice mechanism in a PE learning model. We described potential mechanisms involved in the learning process and compared models implementing these mechanisms in terms of how well they accounted for the learning curve and RT data from a reinforcement-

Table 2 Summary of posterior distributions

	ON			OFF			Contrast			
Parameter	*m*	HDI		*m*	HDI		*m*	HDI		BF
Boundary separation	1.849	1.524	2.178	1.659	1.424	1.882	0.19	–0.05	0.434	16.5
Boundary change	0.008	–0.1	0.111	0.018	–0.04	0.08	–0.01	–0.1	0.08	0.68
Drift rate scaling	3.566	1.886	5.618	1.864	0.721	2.995	1.702	–0.22	3.592	55.3
Nondecision time	0.326	0.233	0.422	0.224	0.162	0.287	0.102	0.03	0.179	227
Learning rate +	0.032	0.003	0.072	0.057	0.008	0.118	–0.02	–0.08	0.031	0.19
Learning rate –	0.023	0	0.08	0.04	0	0.12	–0.02	–0.11	0.046	0.24

Estimated means for off and on medication groups, as well as the contrast for on–off. Values in the HDI columns represent the 95% highest density intervals. Bayes factors (BF) for directional effects are calculated as $i/(1 - i)$, where i is the integral of the posterior distribution from 0 to $+\infty$

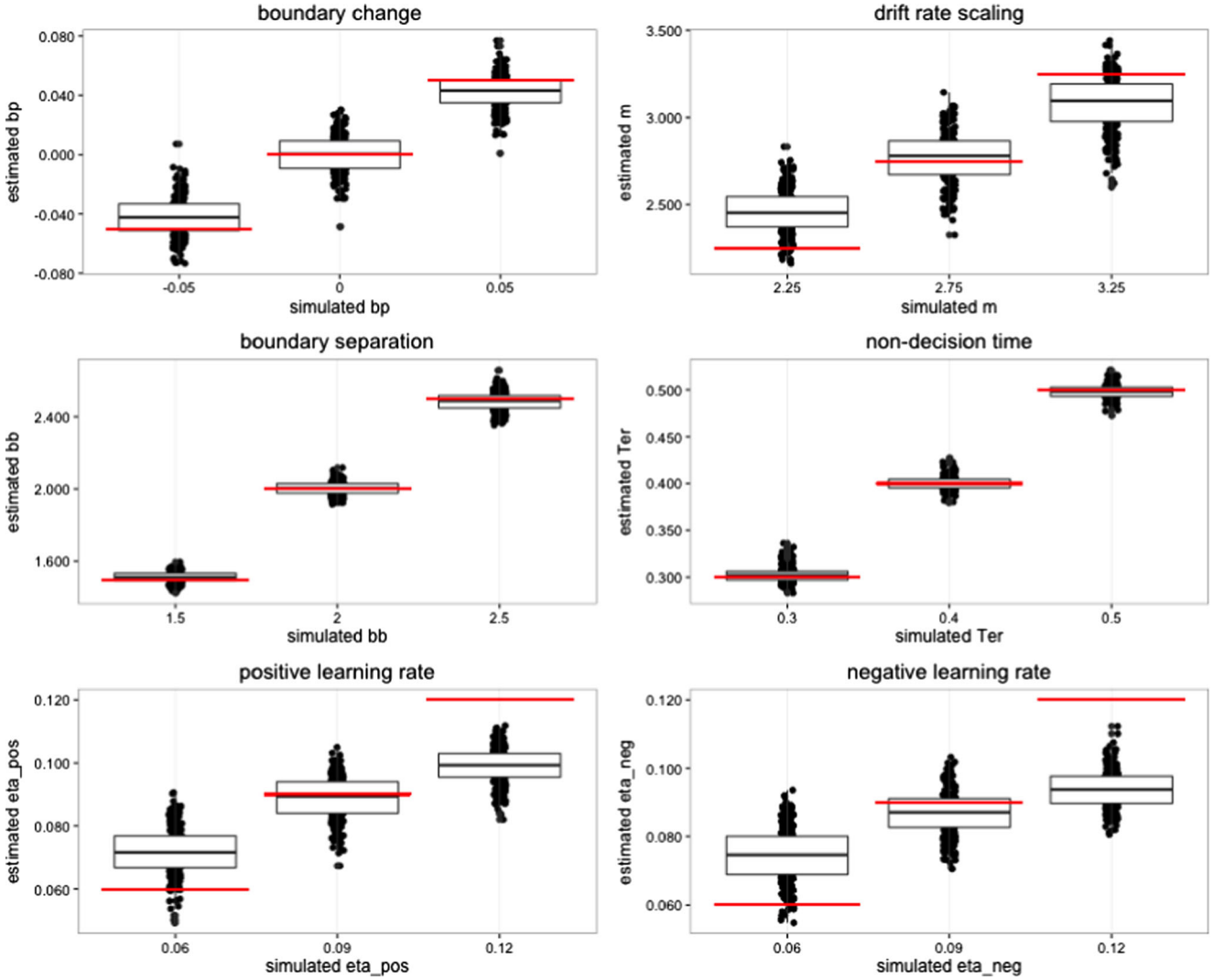

Fig. 6 Parameter recovery results: Mean group parameter values (black dots) for each of the 729 parameter combinations, separated by simulated values. The horizontal color lines represent simulated parameter values, and the boxplots represent distributions of the mean estimated values separated by the simulated parameter values. bb = boundary baseline, bp = boundary power, eta_pos = learning rate for positive prediction errors (PEs), eta_neg = learning rate for negative PEs, m = drift rate scaling, T_{er} = nondecision time

based decision-making task. Using the absolute fit and simulation fit criteria as instantiations of posterior predictive checks, we showed that the best-fitting model accounted for the main choice and RT patterns of the experimental data. The model included independent learning rates for positive and negative PEs, used to update the expected rewards. The differences in expected rewards were scaled by a constant, trial-independent factor to obtain drift rates, and boundary separation was estimated as a trial-dependent parameter. The model was further used to test the effects of stimulant medication in ADHD, and the treatment was found to increase that drift rate scaling and nondecision time, to widen the boundary separation, and to decrease learning rates. Finally, a parameter recovery analysis documented that generative parameters could be estimated in a hierarchical Bayesian modeling approach, thus providing a tool with which one can simultaneously estimate learning and choice dynamics.

Limitations

A number of limitations can be traced to our decision to limit the complexity of our new model. Since this was a first attempt at simultaneously estimating RL and DDM parameters, we did not estimate starting points or include parameters for trial-to-trial variability in decision processes that are included in the full DDM. These parameters are especially useful to capture a number of RT effects, such as differences in the RTs for correct and error responses (Ratcliff & McKoon, 2008). Our modeling did not fully investigate whether learning results in increasing drift rates, or if learning results influences the starting point of the diffusion process. Although the comparison of correct and error

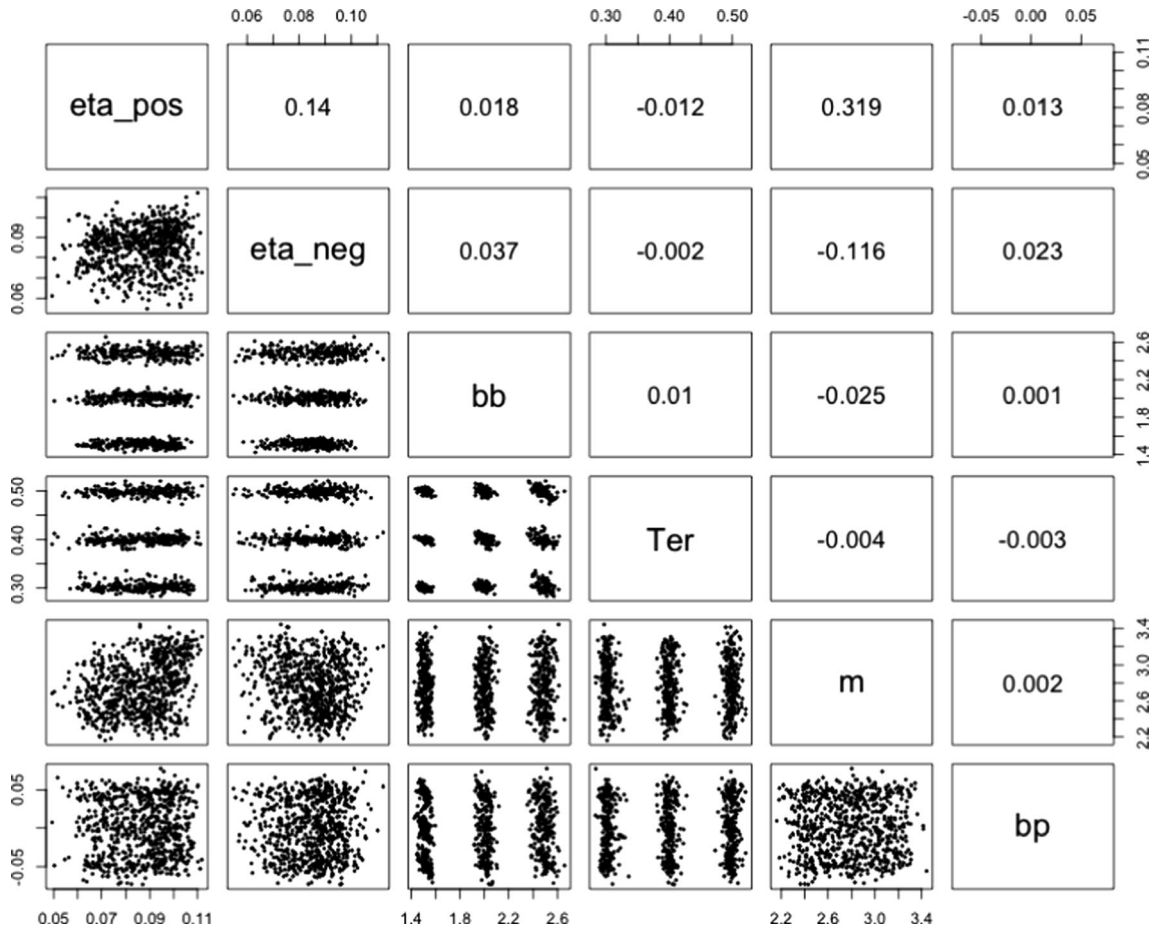

Fig. 7 Scatterplots and correlations between the mean parameter estimates for each of the 64 parameter combination estimations. bb = boundary baseline, bp = boundary power, eta_pos = learning rate for positive prediction errors (PEs), eta_neg = learning rate for negative PEs, m = drift rate scaling, T_{er} = nondecision time

responses revealed very weak evidence for their difference, there was also no clear evidence that they were identical. Hence, further research could explicitly compare models that assume an influence of learning on the drift rate or starting point. In addition, explicitly modeling posterror slowing through wider decision boundaries following errors (Dutilh et al., 2012) could have further improved the model fit. Still, these additional parameters are likely not crucial in the context of the analyzed experiment, because the posterior predictive check showed that the model described the RT distribution data well; for instance, the data did not contain large numbers of slow responses not captured by the model.

A closer examination of the choice data reveals more room to improve the modeling. Even though the learning model is already relatively flexible, it cannot account for all of the choice patterns. The model tended to overestimate learning success for the most difficult learning condition, in which both choice options have similar reward probabilities (60:40). Also, whereas the difference between the proportions of correct responses between learning conditions ON medication was larger at the beginning of learning than at the end, the model predicted larger differences at the end. Such patterns could potentially be captured by models with time-varying learning rates (Frank et al., 2015; Krugel et al., 2009; Nassar, Wilson, Heasly, & Gold, 2010), by models that could explicitly account for capacity-limited and delay-sensitive working memory in learning (Collins & Frank, 2012), or by more elaborate model-based approaches to instrumental learning (Collins & Frank, 2012; Biele, Erev, & Ert, 2009; Doll et al., 2014; Doll, Simon, & Daw, 2012). Hence, whereas the model implements important fundamental aspects of instrumental learning and decision making, the implementation of additional processes might be needed to fully account for the data from other experiments. Note, though, that increasing the model complexity would typically make it more difficult to fit the data, and thus should always be accompanied by parameter recovery studies that test the interpretability of the parameters. As is often seen for RL models, our model did not account for short-time fluctuations in choice behavior. We suggest that reduced short-time fluctuation in simulated choices can be attributed to the fact that the average choice proportions in absolute-fit methods are the result of 100 times as many choices as in the original data, which

effectively reduces variation in the choice proportions between bins. By comparison, the overestimation of learning when the choice options have similar reward probabilities could point to a more systematic failure of the model.

The parameter recovery experiment showed that we were able to recover the parameter values. Although the results showed that we could recover the precise parameter values for the boundary parameters, nondecision time, and drift rate scaling, it proved hard to recover the high and low learning rates, especially for negative PEs. The fact that it is easier to recover positive learning rates is likely due to the fact that there are more trials with positive PEs, as would be expected in any learning experiment. Still, it should be noted that we were able to recover the correct order of the learning rate parameters on the group level. Additional parameter recovery experiments with only one learning rate for positive and negative PEs resulted in more robust recovery of the learning rates, highlighting the often-observed fact that the price of increased model complexity is a less straightforward interpretation of the model parameters (results are available upon request from the authors). In a nutshell, the parameter recovery experiment showed that although we could detect which group had higher learning rates on average, one should not draw strong conclusions on the basis of small differences between learning rates on the individual level.

Effects of stimulant medication on submechanisms of learning and choice mechanisms in ADHD

We investigated the effects of stimulant medication on learning and decision making (Fig. 5), both to compare these results with the observed results from the original article (Frank, Santamaria, et al., 2007) and to assess the RLDD model's ability to decompose choice patterns into underlying cognitive mechanisms. The original article reported selective neuromodulatory effects of dopamine (DA) on go-learning and of noradrenaline (NA) on task switching (Frank, Santamaria, et al., 2007). A comparison of the parameters could therefore describe the underlying mechanisms driving these effects.

Learning rate

The within-subjects effect of stimulant medication identified decreased learning rates for positive and negative feedback following medication. Although it might at first seem surprising that the learning rate was higher off medication, it is important to note that the faster learning associated with higher learning rates also means greater sensitivity to random fluctuations in the payoffs. We found a stronger positive correlation between learning rate and accuracy when patients were on as compared to off medication, selectively for learning rates for positive PEs [interaction effect: $\beta = 0.84$, $t(25) = 2.190$, $p = .038$]. These results show that patients had a more adaptive learning rate on stimulant medication, and also suggest that a reasonably higher scaling parameter for differences in reward expectation is needed to detect the effects of learning rate on learning success.

Drift rate scaling

The drift rate parameter in the RLDD model depends on both learning rate and sensitivity to reward. The drift rate scaling parameter in our model describes the degree to which current knowledge is used, as well as the level of exploration versus exploitation. Stimulant medication was found to increase sensitivity to reward. These results are in line with the involvement of DA in improving the signal-to-noise ratio of cortical representations (Durstewitz, 2006) and striatal filtering of cortical input (Nicola, Hopf, & Hjelmstad, 2004), and in maintaining decision values in working memory (Frank, Santamaria, et al., 2007). They are also supported by the opponent actor learning model, hypothesizing that DA increases sensitivity to rewards during choice, independently from learning (Collins & Frank, 2014).

Boundary separation

Boundary separation estimates increased with medication, indicating a shift toward a stronger focus on accuracy in the speed–accuracy trade-off. This effect is particularly interesting, in that it reveals differences in choice processes during instrumental learning. It also extends the finding of impaired regulation of the speed–accuracy trade-off during decision making in ADHD (Mulder, Bos, Weusten, & van Belle, 2010) to the domain of instrumental learning. The effect can be related to difficulties with inhibiting responses, in line with the dual-pathway hypothesis of ADHD (Sonuga-Barke, 2003), since responses are given before sufficient evidence is accumulated. A possible neural explanation of this effect starts with the recognition that stimulant medication also modulates NA levels (Berridge et al., 2006; Devilbiss & Berridge, 2006), which, via the subthalamic nucleus (STN), provides a global "hold your horses" signal to prevent premature responding (Cavanagh et al., 2011; Frank, 2006; Frank et al., 2015; Frank, Samanta, Moustafa, & Sherman, 2007; Frank, Scheres, & Sherman, 2007; Ratcliff & Frank, 2012).

Nondecision time

Finally, within-subjects contrasts identified a strong increase in nondecision time through medication, which partially (over and above changes in boundary separation) can explain the finding of slower RTs in the medicated group (Frank, Santamaria, et al., 2007). Why stimulant medication should

affect nondecision time is not immediately clear. However, faster nondecision times in ADHD have been reported in studies comparing DDM parameters on unmedicated children with ADHD and in typically developing controls, with an overall effect size of 0.32 (95% CI: 0.48–0.15; Karalunas, Geurts, Konrad, Bender, & Nigg, 2014). The studies reporting this effect could not find a clear interpretation or possible mechanism driving this change, instead suggesting that it might be related to motor preparation and not stimulus encoding (Metin et al., 2013). Alternatively, increased communication with STN through phasic NA activity could also explain how the STN can suppress premature responses (Aron & Poldrack, 2006).

Implications

Modeling choices during instrumental learning with sequential-sampling models could be useful in several ways to better understand adaptive behavior. One topic of increasing interest is response vigor during instrumental learning (see, e.g., Niv, Daw, Joel, & Dayan, 2006). Adaptive learners adjust their response rates according to the expected average reward rate, whereby adaptation is thought to depend on DA signaling (Beierholm et al., 2013). The RLDD model could inform about response vigor adaptations in several ways. For example, average reward expectations in cognitive perceptual tasks can be modeled through PE learning, whereas the adaptation of boundary separation can function as an indicator for the adjustment of response vigor. More generally, the adaptive adjustment of response vigor should result in reduced boundary separations over time in instrumental-learning tasks, as well as (crucially) a greater reduction of boundary separation for decision makers with higher average reward expectations, which would be indicated by a higher drift rate. On a psychological level, the joint consideration of (change of) boundary separation and drift rate can help clarify how the shift from explorative to exploitative choices, fatigue, or boredom influence decision making in instrumental learning. In addition to supporting the exploration of basic RL processes, the RLDD model should also be useful in shedding light on cognitive deficiencies of learning and on decision making in clinical groups (Maia & Frank, 2011; Montague, Dolan, Friston, & Dayan, 2012; Ziegler, Pedersen, Mowinckel, & Biele, 2016), as in the effect of stimulant medication on cognitive processes in ADHD shown here (Fig. 5), but also in other groups with deficient learning and decision making (Mowinckel, Pedersen, Eilertsen, & Biele, 2015), such as in drug addiction (Everitt & Robbins, 2013; Schoenbaum, Roesch, & Stalnaker, 2006), schizophrenia (Doll et al., 2014), and Parkinson's disease (Frank, Samanta, et al., 2007; Moustafa, Sherman, & Frank, 2008; Yechiam, Busemeyer, Stout, & Bechara, 2005).

References

Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32,* 1376–1402. doi:10.1080/03640210802352992

Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R., & Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics, 4,* 95–110. doi:10.1037/a0020684

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: Role of the subthalamic nucleus. *Journal of Neuroscience, 26,* 2424–2433. doi:10.1523/jneurosci.4682-05.2006

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences, 107,* 21767–21772. doi:10.1073/pnas.0908104107/-/DCSupplemental

Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., zel, E. D. U., Dolan, R. J., & Dayan, P. (2013). Dopamine modulates reward-related vigor. *Neuropsychopharmacology, 38,* 1495–1503. doi:10.1038/npp.2013.48

Berridge, C. W., Devilbiss, D. M., Andrzejewski, M. E., Arnsten, A. F. T., Kelley, A. E., Schmeichel, B., … Spencer R. C. (2006). Methylphenidate preferentially increases catecholamine neurotransmission within the prefrontal cortex at low doses that enhance cognitive function. *Biological Psychiatry, 60*(10), 1111–1120. doi:10.1016/j.biopsych.2006.04.022

Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology, 53*(3), 155–167. doi:10.1016/j.jmp.2008.05.006

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113,* 700–765. doi:10.1037/0033-295X.113.4.700

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178. doi:10.1016/j.cogpsych.2007.12.002

Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 538–564. doi:10.1037/0278-7393.11.3.538

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment, 14,* 253–262. doi:10.1037/1040-3590.14.3.253

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459. doi:10.1037/0033-295X.100.3.432

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58,* 313–323.

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General, 143,* 1476–1488. doi:10.1037/a0035813

Cavanagh, J. F., Wiecki, T. V., & Cohen, M. X. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature, 14,* 1462–1467. doi:10.1038/nn.2925

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience, 35,* 1024–1035. doi:10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on

reinforcement learning and choice incentive. *Psychological Review, 121,* 337–366. doi:10.1037/a0037015

Cox, S., Frank, M. J., Larcher, K., Fellows, L. K., & Clark, C. A. (2015). Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. *NeuroImage, 109,* 95–101. doi:10.1016/j.neuroimage.2014.12.070

Craigmile, P. F., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika, 75,* 613–632. doi:10.1007/s11336-010-9172-6

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879. doi:10.1038/nature04766

Devilbiss, D. M., & Berridge, C. W. (2006). Low-dose methylphenidate actions on tonic and phasic locus coeruleus discharge. *Journal of Pharmacology and Experimental Therapeutics, 319,* 1327–1335. doi:10.1124/jpet.106.110015

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology, 22,* 1075–1081. doi:10.1016/j.conb.2012.08.003

Doll, B. B., Waltz, J. A., Cockburn, J., Brown, J. K., Frank, M. J., & Gold, J. M. (2014). Reduced susceptibility to confirmation bias in schizophrenia. *Cognitive, Affective, & Behavioral Neuroscience, 14,* 715–728. doi:10.3758/s13415-014-0250-6

Durstewitz, D. (2006). A few important points about dopamine's role in neural network dynamics. *Pharmacopsychiatry, 39,* 72–75. doi:10.1055/s-2006-931499

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U., & Wagenmakers, E.-J. (2012). How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology, 56,* 208–216. doi:10.1016/j.jmp.2012.04.001

Everitt, B. J., & Robbins, T. W. (2013). From the ventral to the dorsal striatum: Devolving views of their roles in drug addiction. *Neuroscience & Biobehavioral Reviews, 37,* 1946–1954. doi:10.1016/j.neubiorev.2013.02.010

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology, 67,* 641–666. doi:10.1146/annurev-psych-122414-033645

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. D. (2011). The speed–accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience, 31,* 17242–17249. doi:10.1523/JNEUROSCI.0309-11.2011

Forstmann, B. U., & Wagenmakers, E.-J. (Eds.). (2015). *An introduction to model-based cognitive neuroscience*. New York, NY: Springer. doi:10.1007/978-1-4939-2236-9

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks, 19,* 1120–1136. doi:10.1016/j.neunet.2006.03.006

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience, 35,* 485–494. doi:10.1523/JNEUROSCI.2036-14.2015

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences, 104,* 16311–16316. doi:10.1073/pnas.0706111104

Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses: Impulsivity, deep brain stimulation, and medication in Parkinsonism. *Science, 318,* 1309–1312. doi:10.1126/science.1146157

Frank, M. J., Santamaria, A., O'Reilly, R. C., & Willcutt, E. (2007). Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology, 32,* 1583–1599. doi:10.1038/sj.npp.1301278

Frank, M. J., Scheres, A., & Sherman, S. J. (2007). Understanding decision-making deficits in neurological conditions: Insights from models of natural action selection. *Philosophical Transactions of the Royal Society B, 362,* 1641–1654. doi:10.1016/j.braindev.2004.11.009

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science, 306,* 1940–1943. doi:10.1146/annurev.ento.50.071803.130456

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–760.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–472.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review, 22,* 1320–1327. doi:10.3758/s13423-014-0790-3

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, 108*. doi:10.1073/pnas.1014269108

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences, 108,* 18120–18125. doi:10.1073/pnas.1109322108/-/DCSupplemental

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). New York, NY: Springer. doi:10.1007/978-1-4939-2236-9_2

Jeffreys, H. (1998). *The theory of probability*. Oxford, UK: Oxford University Press.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review, 121,* 1–32. doi:10.1037/a0034190

Karalunas, S. L., Geurts, H. M., Konrad, K., Bender, S., & Nigg, J. T. (2014). Reaction time variability in ADHD and autism spectrum disorders: Measurement and mechanisms of a proposed trans-diagnostic phenotype. *Journal of Child Psychology and Psychiatry, 55,* 685–710. doi:10.1111/jcpp.12217

Kayser, A. S., Buchsbaum, B. R., Erickson, D. T., & D'Esposito, M. (2010). The functional anatomy of a perceptual decision in the human brain. *Journal of Neurophysiology, 103,* 1179–1194. doi:10.1152/jn.00364.2009

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences, 108,* 13852–13857. doi:10.1073/pnas.1101328108/-/DCSupplemental

Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S. C., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences, 106,* 17951–17956.

Kruschke, J. K. (2010). *Doing Bayesian data analysis*. San Diego, CA: Academic Press.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling*. Cambridge, UK: Cambridge University Press.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York; NY: Wiley.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Publishing Group, 14,* 154–162. doi:10.1038/nn.2723

Marsman, M., & Wagenmakers, E.-J. (2016). Three insights from a Bayesian interpretation of the one-sided *p* value. *Educational and Psychological Measurement*. doi:10.1177/0013164416669201. Advance online publication.

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. J. S. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology, 27,* 193–200. doi:10.1037/a0031533

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 16,* 1936–1947.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences, 16,* 72–80. doi:10.1016/j.tics.2011.11.018

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs (R package version 0.9.11-1) [Computer software manual]. Retrieved from http://bayesfactorpcl.r-forge.r-project.org

Moustafa, A. A., Sherman, S. J., & Frank, M. J. (2008). A dopaminergic basis for working memory, learning and attentional shifting in Parkinsonism. *Neuropsychologia, 46,* 3144–3156. doi:10.1016/j.neuropsychologia.2008.07.011

Mowinckel, A. M., Pedersen, M. L., Eilertsen, E., & Biele, G. (2015). A meta-analysis of decision-making and attention in adults with ADHD. *Journal of Attention Disorders, 19,* 355–367. doi:10.1177/1087054714558872

Mulder, M. J., Bos, D., Weusten, J. M. H., & van Belle, J. (2010). Basic impairments in regulating the speed–accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological Psychiatry, 68,* 1114–1119. doi:10.1016/j.biopsych.2010.07.031

Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences—A model-based review. *Neuroscience, 277,* 872–884. doi:10.1016/j.neuroscience.2014.07.031

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience, 32,* 2335–2343. doi:10.1523/JNEUROSCI.4156-11.2012

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences, 11,* 49–54. doi:10.1016/j.cobeha.2016.04.003

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience, 30,* 12366–12378. doi:10.1523/JNEUROSCI.0822-10.2010

Nicola, S. M., Hopf, F. W., & Hjelmstad, G. O. (2004). Contrast enhancement: A physiological effect of striatal dopamine? *Cell and Tissue Research, 318,* 93–106. doi:10.1007/s00441-004-0929-z

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2006). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology, 191,* 507–520. doi:10.1007/s00213-006-0502-4

Nunez, M. D., Srinivasan, R., & Vandekerckhove, J. (2015). Individual differences in attention influence perceptual decision making. *Frontiers in Psychology, 8,* 18. doi:10.3389/fpsyg.2015.00018

Pedersen, M. L., Endestad, T., & Biele, G. (2015). Evidence accumulation and choice maintenance are dissociated in human perceptual decision making. *PLoS ONE*, *10*, e140361. doi:10.1371/journal.pone.0140361

Peruggia, M., Van Zandt, T., & Chen, M. (2002). Was it a car or a cat I saw? An analysis of response times for word recognition. In C. Gatsonis (Ed.), *Case studies in Bayesian statistics* (Vol. 6, pp. 319–334). New York, NY: Springer.

Plummer, M. M. (2004). JAGS: Just another Gibbs sampler [Software]. Retrieved from https://sourceforge.net/projects/mcmc-jags/

Plummer, M. M., & Stukalov, A. (2013). Package "rjags." [Software update for JAGS]. Retrieved from https://sourceforge.net/projects/mcmc-jags/

R Development Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R., Cherian, A., & Segraves, M. A. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *Journal of Neurophysiology, 90,* 1392–1407. doi:10.1152/jn.01049.2002

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Computation, 24,* 1186–1229.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356. doi:10.1111/1467-9280.00067

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108,* 370–392. doi:10.1037/0033-295X.108.2.370

Schoenbaum, G., Roesch, M. R., & Stalnaker, T. A. (2006). Orbitofrontal cortex, decision-making and drug addiction. *Trends in Neurosciences, 29,* 116–124. doi:10.1016/j.tins.2005.12.006

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.

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

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences, 27,* 161–168. doi:10.1016/j.tins.2004.01.006

Solomon, M., Frank, M. J., & Ragland, J. D. (2015). Feedback-driven trial-by-trial learning in autism spectrum disorders. *American Journal of Psychiatry, 172,* 173–181. doi:10.1176/appi.ajp.2014.14010036

Sonuga-Barke, E. J. S. (2003). The dual pathway model of AD/HD: an elaboration of neuro-developmental characteristics. *Neuroscience & Biobehavioral Reviews, 27*, 593–604. doi:10.1016/j.neubiorev.2003.08.005

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision, 1,* 161–183. doi:10.1037/dec0000005

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review, 122,* 312–336. doi:10.1037/a0038894

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550–592. doi:10.1037/0033-295X.111.3.757

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16,* 44–62. doi:10.1037/a0021765

Vehtari, A., Gelman, A., & Gabry, J. (2016). *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC*. Retrieved from https://arXiv.org/abs/1507.04544

Wabersich, D., & Vandekerckhove, J. (2013). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods, 46,* 15–28. doi:10.3758/s13428-013-0369-3

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener package: An R package providing distribution functions for the Wiener diffusion model. *R Journal, 6,* 49–56.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology, 54,* 14–27. doi:10.1016/j.jmp.2008.12.001

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review, 19,* 1057–1064. doi:10.3758/s13423-012-0295-x

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54,* 39–52. doi:10.1016/j.jmp.2010.01.004

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics, 7,* 14. doi:10.3389/fninf.2013.00014

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science, 16,* 973–978. doi:10.1111/j.1467-9280.2005.01646.x

Ziegler, S., Pedersen, M. L., Mowinckel, A. M., & Biele, G. (2016). Modelling ADHD: a review of ADHD theories through their predictions for computational models of decision-making and reinforcement learning. *Neuroscience & Biobehavioral Reviews*. doi:10.1016/j.neubiorev.2016.09.002