# Speed and accuracy in learning: A combined Q-learning diffusion decision model analysis

**Steven Miletić (s.miletic@uva.nl)**
IMCN research unit, Department of Psychology, University of Amsterdam
Nieuwe Achtergracht 129B, 1001 NK Amsterdam, The Netherlands

**Russell J. Boag (r.j.boag@uva.nl)**
IMCN research unit, Department of Psychology, University of Amsterdam
Nieuwe Achtergracht 129B, 1001 NK Amsterdam, The Netherlands

**Varvara Mathiopoulou (mathiopoulou.barbara@gmail.com)**
IMCN research unit, Department of Psychology, University of Amsterdam
Nieuwe Achtergracht 129B, 1001 NK Amsterdam, The Netherlands

**Birte U. Forstmann (buforstmann@gmail.com)**
IMCN research unit, Department of Psychology, University of Amsterdam
Nieuwe Achtergracht 129B, 1001 NK Amsterdam, The Netherlands

**Abstract:**

**Recent advances in cognitive modelling merged two classes of cognitive models: Sequential sampling models of decision-making, and reinforcement learning models of error-driven learning. Such integrated models provide theoretical accounts of the cognitive processes underlying decisions and learning simultaneously. Here, we test whether a classical decision-making phenomenon -the speed-accuracy trade-off- can be observed in an instrumental learning task, and whether an integrated reinforcement learning/sequential sampling model is able to capture this effect. The results show that the model indeed captured the speed-accuracy trade-off effect in empirical data, as well as changes in response times and accuracies due to learning over the course of the experiment. This study further illustrates the great promise of the integration of the reinforcement learning and sequential sampling frameworks for cognitive psychology and cognitive neuroscience.**

**Keywords: Q-learning; sequential sampling models; diffusion decision model; speed-accuracy trade-off**

## Introduction

Recent advances in cognitive modelling (Fontanesi, Gluth, Spektor, & Rieskamp, 2019; Pedersen, Frank, & Biele, 2017; Sewell, Jach, Boag, & Van Heer, 2019) shows an increasing interest in the merger of two prominent classes of cognitive models: Sequential sampling models (SSMs) of decision-making (Forstmann, Ratcliff, & Wagenmakers, 2016), and reinforcement learning (RL) models of model-free, error-driven learning (Sutton & Barto, 2018). Such a merger holds the promise of the best of both worlds: The SSM provides a mechanistic theory of the cognitive processes underlying decisions, while the RL model accounts for learning from errors over time. Merged RL-SSM models have now been shown to be able to explain response time distributions of decisions as well as increases in speed and accuracy over the course of an experiment due to learning (Fontanesi et al., 2019; Pedersen et al., 2017; Sewell et al., 2019).

The use of an SSM in learning tasks as the choice function implies that the cognitive processes underlying decisions in learning tasks are identical (or very similar) to the processes more often studied in perceptual decision-making. If this is the case, it can be expected that well-studied decision-making phenomena are also present in learning tasks. In support of this hypothesis, Pederson et al. (2017), Sewell et al. (2019), and Fontanesi et al. (2019) found typical choice difficulty effects, and Fontanesi et al. (2019) furthermore found magnitude effects in a learning task. Here, we extend these findings by testing for the presence of a speed-accuracy trade-off in the decision-making process during a learning task.

The speed-accuracy trade-off (SAT) refers to the long-studied phenomenon that people are able to voluntarily trade off decision speed for decision accuracy (Heitz, 2014). In the field of perceptual decision-making, this ability is often targeted in experimental manipulations to study decision-making behavior under speed stress, as well as the neural underpinnings of responses (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010). SSMs provide an intuitive mechanistic account of the SAT. These models assume that people make decisions by gradually accumulating noisy evidence until a threshold level of evidence is reached, at which point people commit to a decision and initiate a motor response. SAT settings are captured by the threshold parameter: Increases in threshold lead to slower but more accurate decisions, and vice versa.

Accordingly, we hypothesized that the threshold parameter is affected by SAT instructions in a learning task as well.

Furthermore, we explore whether the learning rate parameter is also affected. In what follows, we test for the presence of a SAT in decision-making in a typical learning task.

## Methods

### Data collection

**Participants** 35 healthy participants (age 20.5y [SD 2.5y], 8 male) were recruited from the subject pool of the department of Psychology, University of Amsterdam. The study was approved by the local ethical committee. All participants gave written informed consent and received course credits for participating. Participants could earn extra course credit by earning many points in the task.

**Task** The task used was an instrumental learning task (c.f. Sewell et al., 2019; Figure 1), in which participants need to learn by trial and error, which of two visually presented choice options are mostly likely to lead to a reward. Within a block of this task, three stimulus pairs were presented. Each pair had fixed reward probabilities associated with each stimulus. Trials start with the presentation of two abstract symbols (characters in the Agathodaemon alphabet), between which the participant has to choose. The choice is highlighted, after which the participant receives feedback about the outcome of the choice (0 or 100 points), and the actual reward. The actual reward was equal to the outcome of the choice if the participant chose in time, or a penalty of -100 (irrespective of the choice outcome) if the participant was too late.

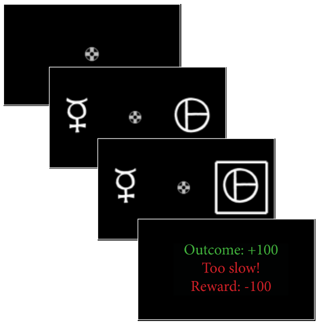

Figure 1: Illustration of the task. The feedback shown here is one of three possible feedback screens. In this case, the participant chose a symbol that led to +100 points outcome, but was too slow, and therefore lost 100 points. The other feedback screens were for choices in time, where the reward was equal to the choice outcome.

**Design** The experiment consisted of three blocks. The first block was a calibration block, in which the participant had to learn four stimuli sets (with reward probabilities of 80%-20%; 70%-30%; 65%-35%; and 60%-40%). The data of this block were used to determine the difficulty of the upcoming blocks (an accuracy-weighted sum of the reward probabilities of each stimulus set in the calibration block), as well as to set an individual *response deadline* in trials in which participants were instructed to respond *fast* (the 65th quantile of the RT distributions in the calibration block, plus a random intercept sampled from an exponential distribution to reduce deadline predictability).

Subsequently, participants performed two more blocks, in which each trial was preceded by a speed or accuracy cue. In speed trials, participants were required to respond before the deadline, and too late responses were never rewarded. Each block consisted of 304 trials (152 trials per cue, 76 trials per stimuli pair, 4 stimuli pairs). In one of the two blocks, speed and accuracy trials were randomly intermixed. In the other block, the cues were presented in miniblocks of 8 trials. Cue order was randomized across participants. Here, we only analyze the data of the miniblocks, since we found the behavioral effect of the cues to be largest in this block.

### Cognitive modelling

The data were modelled using a combination of Q-learning (Sutton & Barto, 2018) and the diffusion decision model (DDM; Ratcliff, 1978). Like SSMs, we assume that participants make decisions by gradually accumulating evidence over time (with a mean accumulation rate called the *drift rate*) until a threshold level of evidence is reached, and a decision is made. However, whereas the drift rate is estimated as a free variable in standard SSMs, we assume that the drift rate is a linear function of the difference in Q-values. Q-values can be interpreted as the reward expectations for each *state-action pair (s, a)*, and are updated after every trial *t* according to:

$$Q_{(s,a)_{t+1}} = Q_{(s,a)_t} + \alpha(r_t - \max_a Q_{(s,a)_t})$$

where $0 < \alpha \leq 1$ is the learning rate, and *r* reward. In all the models we fit, the DDM had four parameters: a drift rate *v*, threshold *b*, and two parameters describing a uniform non-decision time distribution $t_0$ and $st_0$. Non-decision time variability was included because it improved the quality of fit. We assumed that drift rate $v$ on every trial was a function of the difference in Q-values for both choice options, linearly scaled by a factor $m$:

$$v_t = m\left(Q_{(s,a_1)_t} - Q_{(s,a_2)_t}\right)$$

We fit four different model specifications (Table 2), allowing the learning rate, the threshold, or both parameters to vary with the speed-accuracy trade-off instruction. Models were fit using maximum likelihood estimation. For formal model comparison, we computed the Bayesian information criterion (BIC; Wagenmakers & Farrell, 2004). The BIC is defined as $BIC = -2\log(L) + k\log(n)$, where $k$ is the number of

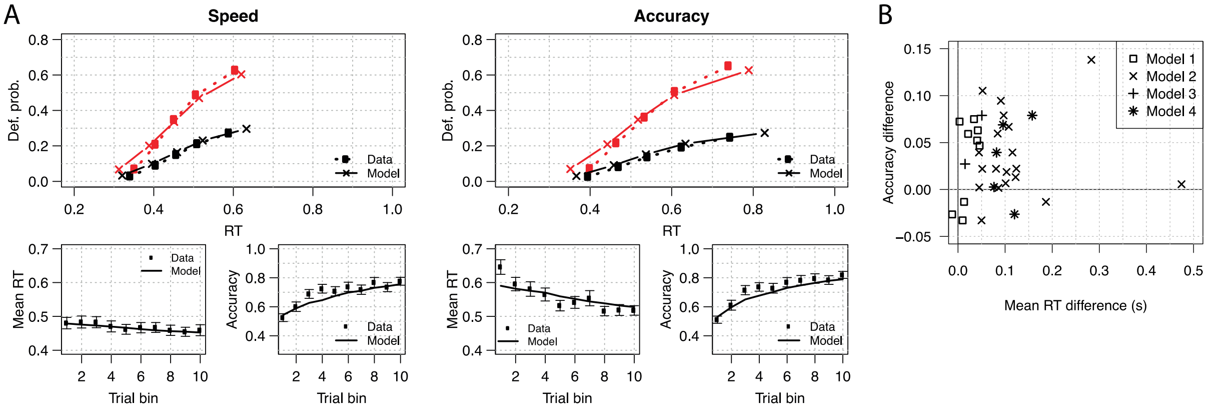

Figure 2. A) Quality of fit, averaged over participants, for both the speed (left panels) and accuracy (right panels) conditions. Upper panels are defective cumulative density functions of data (dotted lines, circles) and model prediction (solid lines, crosses). Colors indicate the choice option (red is the optimal, "correct" choice). Lower panels show the changes over the course of the experiment in mean RT and accuracy. Trials were binned into groups of 30 (10 bins total). Error bars are the standard error. B) Relationship between cue-induced changes in mean RT, in accuracy, and the winning models.

estimated parameters, and *n* the number of observations (trials). Lower values indicate better trade-offs between quality of fit and model complexity.

## Results

**Descriptives** In the *speed* condition, choices were faster (*speed* M = 0.467s [SD = 0.08s], *accuracy* M = 0.556s [SD = 0.08s], $t(34) = -6.024$, $p < 0.001$) and less accurate (*speed* M = 0.697 [SD = 0.138], *accuracy* M = 0.724 [SD = 0.135], $t(34) = -3.703$, $p < 0.001$) than in the *accuracy* condition, in line with a typical SAT effect.

**Model comparison** Table 1 provides an overview of the model comparison. The overall winner was Model 2, with the lowest summed BIC and highest number of participants for which it was preferred. The model with separate thresholds for speed and accuracy trials won for 19 participants (54% of all participants). For 9 participants (26%), there was no evidence for a change in threshold or learning rate; for another 5 subjects (14%), there was evidence for both a change in threshold and learning rate, and the final 2 participants (5%) showed evidence for a change in learning rate only.

Table 1: Model comparison. M = model number, *k* = number of free parameters, $\sum BIC$ = summed BIC (lower is better), *n* = number of participants for which the model won.

| M | Free parameters | *k* | $\sum BIC$ | *n* |
|---|---|---|---|---|
| 1 | $b$, $\alpha$, $t_0$, $st_0$, $m$ | 5 | -1350.83 | 9 |
| 2 | $b_{spd}$, $b_{acc}$, $\alpha$, $t_0$, $st_0$, $m$ | 6 | -2265.55 | 19 |
| 3 | $b$, $\alpha_{spd}$, $\alpha_{acc}$, $t_0$, $st_0$, $m$ | 6 | -1222.59 | 2 |
| 4 | $b_{spd}$, $b_{acc}$, $\alpha_{spd}$, $\alpha_{acc}$, $t_0$, $st_0$, $m$ | 7 | -2129.82 | 5 |

**Model fits** Figure 2A illustrates the overall quality of fit for both conditions for Model 2. The upper panels show the defective cumulative density functions of both the data and the model predictions (mean across participants and model predictions). The mean RTs and accuracies are captured well by the model, although the model predicts a slightly more skewed distribution (as indicated by an overestimated defective probability in the left tail, and an underestimation in the right tail). We briefly return to this in the Discussion. The lower panels show the changes in response time and accuracy over the course of the experiment due to learning. The model captures the trends in both dependent variables well.

**Manipulation effect size and model preference** It is interesting to explore why no effect was found for a quarter of the participants. One potential reason is that the behavioral effect was too small (e.g. compared to the effects reported in Forstmann et al., 2008) to be detected using the BIC metric for model comparisons, which is known to be relatively conservative (Wagenmakers & Farrell, 2004). To test this hypothesis, we calculated for each participant the difference (between speed and accuracy cues) in mean RT and accuracy. Figure 2B illustrates the winning model for each combination of difference in RT and accuracy between SAT conditions. For all participants with an RT difference of roughly 50ms or larger, the winning model always included a threshold change. For all participants with a smaller RT difference, the winning model never included a threshold change. This suggests that for the subset of 9 participants for which the null model won, the effect size of the experimental manipulation was not sufficiently large to warrant an extra parameter in the model.

## Discussion

Here, we tested whether the speed-accuracy trade-off, a hallmark-phenomenon in perceptual decision-making, can be observed in an instrumental learning task. In line with expectations, instructing participants to respond quickly led to faster but less accurate decisions than instructing them to be accurate. An RL-DDM was able to capture the response time distributions, response accuracy, as well changes in response time and accuracy over time due to learning. Model comparisons revealed that the behavioral changes due to speed and accuracy instructions were caused by changes in threshold settings, with little evidence for changes in learning rates.

Several interesting aspects should be noted. Firstly, the effects of the SAT manipulations on response times and accuracies were small compared to those commonly observed in perceptual decision-making tasks (e.g., Forstmann et al., 2008). One potential reason may be that the incentive to be accurate was always stronger than to be fast, because accuracy was explicitly rewarded with course credit at the end of the experiment. In future studies, it may be possible to increase the effect size by providing an extra incentive to be fast after speed cues, for example by rewarding response speed.

Secondly, there is some misfit in the response time distributions, mostly with respect to the skewness: the RL-DDM consistently predicted more right-skewed distributions (skewness >1.5) than observed in the data (median skewness = 1.0 [IQR 0.64]). The inclusion of the non-decision time variability reduced the misfit to some extent. Including non-decision time variability in the model may be especially important to include in this task due to the high response speeds (compared to perceptual decision-making tasks). As such, non-decision time variability has a relatively large influence on the response time distributions. However, even with the inclusion of non-decision time variability, the models still predict higher skewness. Future research should focus on testing which cognitive processes (e.g., urgency; Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015) can explain this feature in the data.

Thirdly, and advocating the integration of models from the RL and SSM frameworks, our conclusions with respect to the effect of the SAT manipulation could not have been drawn based on classical RL modelling using softmax as a choice function. Fitting RL models with softmax to the same data led to the overall conclusion that neither the learning rate nor softmax' inverse temperature parameter was affected by the SAT manipulation, even though there was clearly a behavioral effect in both accuracy and RT. The combined model was able to capture both the SAT effect as well as the learning effect in our data. This illustrates the theoretical advantage of combining insights from the RL and SSM frameworks, as well as the great promise of this integration for cognitive psychology and cognitive neuroscience.

## Acknowledgments

This work was supported by an NWO-VICI grant (BUF).

## References

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, *33*(1), 10–16.

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*.

Forstmann, B. U., Dutilh, G., Brown, S. D., Neumann, J., Cramon, D. Y. Von, & Ridderinkhof, K. R. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(45), 17538–17542.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*(1), 641–666.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the Evidence for Collapsing Boundaries and Urgency Signals in Perceptual Decision-Making. *Journal of Neuroscience*, *35*(6), 2476–2484.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*(8 JUN), 1–19.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin and Review*, *24*(4), 1234–1251.

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

Sewell, D. K., Jach, H. K., Boag, R. J., & Van Heer, C. A. (2019). Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychonomic Bulletin and Review*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. *MIT Press* (2nd ed.). Cambridge, MA: MIT press.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.