MBCS Internship Report

Modelling Environmental Volatility in Decision Making

Original title at research proposal: Improvement of RL-EAMs to account for behavioral data

Matheus Boger
9/8/2023

Table of Contents

Abstract

Cognitive modeling is an important tool to investigate behavior and the brain. To this end, two large classes of models have been developed in the past: reinforcement learning (RL) models to portray the changes in behavior as a response to its consequences, and evidence accumulator models (EAMs) to portray the underlying processes of decision-making, including the time it takes to reach a decision. However, the combination of the two methodologies into one is recent, and few is known on how more complex RL learning rules that incorporate environmental volatility in their definitions could be used along EAMs. In this study, the volatile Kalman filter (VKF), as proposed by Piray and Daw, 2020, is tested as such a possible learning rule on a reversal learning task conducted by Miletić et al., 2021. A parameter recovery analysis was also conducted in order to ascertain the robustness of the VKF itself. Results indicated that the VKF has poor recovery of its parameters, especially the ones regarding the modeling of volatility, and that the simplest integration possible of it to an LBA is not enough to make it a good descriptor of participant accuracy and reaction time (RT) data. To solve this issue, it is recommended that future work focus on how the VKF can be improved to be more discrete in nature, or that a more complex linkage to an EAM with an urgency signal be attempted. Nevertheless, the increase of model complexity brought by the VKF was demonstrated to not worsen the BIC of the overall model, which gives hope to the overarching class of multi-layered learning rules.

Introduction

Cognitive modelling is an essential tool in the cognitive scientist's toolkit in the present days (Forstmann et al., 2016). Its aim is to clarify and quantify the relation between observed behavior and latent psychological constructs. Hopefully, the same constructs can then be tied to brain activation patterns in order to establish a causal, or at least correlational, connection between brain and behavior (Palestro et al., 2018).

In the present study, focus is given to two classes of models: reinforcement learning (RL) models (Sutton & Barto, 1998) and evidence accumulator models (EAM) (Donkin & Brown, 2018). RL is based on how learning takes place in an agent-environment interaction. An agent implemented with RL techniques has a set goal and a task-dependent manner to extract information from its environment in order to decide which action should be taken next in order to achieve such goal. As a result of the agent's action, the environment may or may not reward or punish it, and this feedback is then processed as an error signal between what the agent expected to happen and the observed result. From this signal, adjustments on the agent's decision-making process are made, and in this way, it learns how to better achieve its goal through iterations of the algorithm. On the other hand, EAMs are models of decision-making processes themselves. Their core idea is that evidence favoring one or another percept is accumulated through time within a noisy process that reflects the uncertainty embedded in the environment and the agent's cognitive systems. Furthermore, once a certain percept's threshold of evidence is surpassed, that percept is taken to be the final result of the process. If an EAM is used to model sensorial processing, then this percept could represent different stimuli, while if it is used to model decision-making, this percept could represent a behavioral output that the agent ought to do in response to determined stimuli. In any case, the inherent temporal nature of an EAM makes this class of models suitable to fit to reaction time (RT) experiments in cognitive psychology.

As it is possible to note, there is a very convenient linking hook between the two types of models: decision-making. While RL models generally assume a simplified decision-making step for their agent's behavior and focus more on the learning aspect, EAMs detail the decision-making process without taking learning into consideration. The precise mathematical formulation that binds these two classes of models when they are deployed in tandem depends on the exact subtypes of RL models and EAMs used. However, the gist is that an EAM is embedded into an RL paradigm as its decision-making algorithm. This allows researchers to gain insight into both types of processes and their interaction at the same time. As an example, one could look more specifically at the effect that repeated trials of the same task have on participant's RT and accuracy distributions. A possible next step could be to use this mathematical basis to link psychological behavior to neural activation.

This kind of combination of two classes of models into one is called combined modelling (Miletić et al., 2020), and in the current case the final model can be called an RL-EAM. One study that looks at how RL-EAMs can be applied for generating new knowledge is Miletić et al., 2021. In their study, Miletić et al. fitted a series of RL-EAMs to the same observed results of a reversal learning task to further specify which one could account better for the data. The current study aims to test novel RL-EAM techniques that would allow modeling of behavior under volatile environments and is an expansion on their work. Thus, it is crucial to outline aspects of their methods and findings which are further expanded in this study.

A schematic of the task the participants were asked to do can be seen in Figure 1. Participants were shown a fixation cross at the beginning of each trial, which was then followed by the presentation of two stimuli. Each stimulus is stochastically associated to a reward in points with reward probabilities uninformed to the participant, and pairs of stimuli are bound together in a set that is always shown simultaneously, with only the left or right placement of the stimuli being altered. The idea is that over trials, participants will choose either stimulus from each set and will develop a sense of how probable they

are to receive a reward when choosing one or the other. This part of the experiment is called the acquisition phase. At some point midway through the experiment, however, the contingencies of reward switch between the two stimuli that belong to the same set. Again, neither the presence nor the timing of the reversal is informed to the participant (although since every block contains a reversal at a similar point it could be expected that the participants also learn to expect a reversal). At the consequent phase, called reversal phase, the focus of observation is on how quickly both participants and then models adapt to the new condition, changing their behavior accordingly to earn as many points as possible.

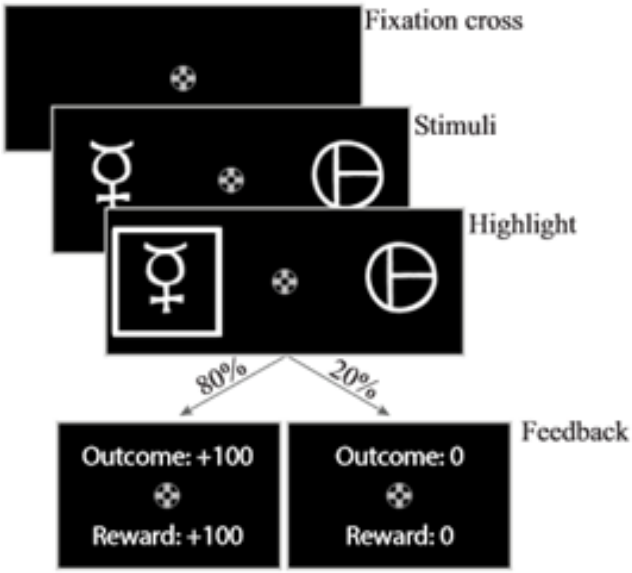

Figure 1. (Adapted from Miletić et al., 2021) Schematic of the experiment conducted by Miletić et al. The highlight step indicates a choice made by the participant, while the percentages indicate the probability of reward associated to each stimulus. The presented feedback screen (positive or negative outcome) depends on the stochastic nature of the reward. It is important to note that midway through a block the probabilities of reward switch between stimuli, so in this case the symbol drawn on the left would then be associated to a 20% chance of reward when chosen, while the one on the right would be associated to an 80% chance.

This task affords the study of RL-EAM modelling because it involves both learning effects and decision-making effects to be taken into consideration. Furthermore, it imposes a larger requirement in the models because they must be able to account for the reversal in stimulus-reward contingencies.

As for the models themselves, Miletić et al. compared 4 different types of RL-EAMs, all differing exclusively on the EAM component. This means that the learning rule used to update the model's behavior in the RL component was kept constant to a simple delta rule. On the other hand, the current study focuses on the improvements that could be achieved if a more complex learning rule was adopted, so one is advised to refer to Miletić et al., 2021 for a detailed account of their models if interested.

Finally, their results. Of relevance to the current study is the fact that their preferred model had two salient misfit points regarding the task. The first one is that RT and accuracy were underestimated during the acquisition phase of the experiment. This misfit, however, is not directly addressed in the current study. The second one is that participants exhibited a faster adaptation to the new contingencies than the model predicts. In other words, the model needed more trials of learning than the participants to be able to adjust its output behavior to be optimal again after a reversal of reward probabilities. This points to a possibility that more information from the environment must be processed in a meaningful way in order

to achieve a similar behavioral result to humans, and so that the model should be altered to bridge this gap.

Suitably, the concept of environmental volatility comes in handy as a starting point for this investigation (Piray & Daw, 2020). Environmental volatility refers to the degree an environment changes in a relevant way to the task at hand during the length of such task. In the case of the reversal learning task, the environmental volatility is taken to be the reversal itself and attributes such as when does it happen or how often it happens. However, the concept in its broad form is also of interest to a cognitive scientist, since the real world is the utmost volatile environment (Behrens et al., 2007). Thus, a study on how decision-making operates under more or less volatile situations is of interest to the study of the mind, and one practical way to start it is through the reversal learning task. If we could fit a model that captures the volatility present in this task well, we could later hope to expand this knowledge to more complex situations or broader tasks. Again, the mathematical underpinning of how volatility is processed could also then be explored to validate whether a similar process indeed happens in the brain. It is concerning, then, that the model deemed best by Miletić et al. failed to grasp the reversal as well as the human participants.

Fortunately, there is literature on the field of RL suggesting more convolute learning rules that take volatility in consideration (Gallistel et al., 2014; Mathys et al., 2011; Piray & Daw, 2020). In the present study, the volatile Kalman filter (VKF) as proposed by Piray and Daw, 2020 is taken as an exploratory alternative for the simple delta rule in the reversal learning task. The VKF is an algorithm that assumes a 2-fold stochastic generative process for the given data and estimates an optimal learning rate for updating the expected outcome of said generative process taking into consideration the volatility, uncertainty and observational noise attributed to the process. Detailed explanation and mathematical definitions of the VKF follow in the Methods section. Here it is important to note that the VKF has a broader range of applications than just the reversal learning task. Its universal structure allows for modelling of learning under diverse volatile environments, and to test it under the reversal learning task is a first step in ascertaining its usefulness or the universality of its design. Thus, the current study not only combines an RL paradigm that uses the VKF as a learning rule to an EAM to compare it against the simple delta rule in the reversal learning task, but it also aims to validate the VKF as a method on itself through a parameter recovery analysis.

To recapitulate: Miletić et al., 2021 applied RL-EAMs to the reversal learning task to study learning and decision-making effects at the same time. Nevertheless, their models failed to represent the speed with which human participants adapted to the reversal. It was posed that this might be due to the learning rule of their model not taking the volatility of the environment into consideration. Then, for the present study, it was hypothesized that such an improvement could be achieved through the implementation of the VKF as a learning rule for the RL-EAM. Thus, this study attempts to answer whether the VKF is a robust method and if it is suitable for application to the reversal learning task.

Methods

This section is subdivided into three segments. First, the tested models are defined. Then, the methodology adopted for the parameter recovery analysis is laid out. Finally, the model fitting and comparison criteria are specified. All the code for the implementation of the methods can be found in the online repository available at http://www.github.com/mathboger/imcn-uva-internship-2023 .

Models

For the current study, three models were implemented (Table 1). They can be classified according to which technique they use for the learning rule and the decision-making process. The simplest model is the baseline model, and it uses a simple delta rule (Equation 1) as its learning rule, and a softmax function (Equation 2) as its decision-making component. The second model is an improvement onto the baseline model that replaces its simple delta rule for the VKF (Equations 3 to 8). This second model was named stepping stone model, since it is in between the other two models in complexity. Finally, the last model is the target model: a full-fledged RL-EAM that makes use of a linear ballistic accumulator (LBA; Equations 9 to 11; a type of EAM) instead of a simple softmax function. An LBA was chosen for the EAM in the current study because it is the most used and well supported EAM, and thus easy to embed into a self-implemented RL paradigm with the VKF. Furthermore, the simple delta rule RL-LBA was omitted from the current study due to time constraints and to similar simple EAMs having already been tested by Miletić et al. in their study.

Table 1. Models implemented for the current study (number of total parameters in parentheses).

Learning rule / Decision-making	Softmax	LBA
Simple delta rule	Baseline model (2)	Omitted
VKF	Stepping stone model (5)	Target model (8)

Equation 1. Simple delta rule

$$(1)\ Q_t = Q_{t-1} + \alpha(o_t - Q_{t-1})$$

Where Q_t is the Q-value associated to a stimulus at trial t, α is the learning rate and o_t is the observation made (reward obtained) at trial t. This poses that the values associated to a stimulus only depend on the previous value, updated by a prediction error.

Equation 2. Softmax function

$$(2)\ \sigma(\boldsymbol{Q})_i = \frac{e^{\beta Q_i}}{\sum_{j=1}^{K} e^{\beta Q_j}}$$

Where $\sigma()_i$ represents the softmax function output on the i-th stimulus, **Q** is the vector of Q-values associated to all K stimuli taken into consideration and β is the inverse temperature parameter, which determines how much bias should be put into stimuli with higher Q-values.

Equations 3 to 8. Volatile Kalman filter

$$(3)\ k_t = \frac{w_{t-1} + v_{t-1}}{w_{t-1} + v_{t-1} + \omega}$$

$$(4)\ \alpha_t = \sqrt{w_{t-1} + v_{t-1}}$$

$$(5)\ m_t = m_{t-1} + \alpha(o_t - s(m_{t-1}))$$

$$(6)\ w_t = (1 - k_t)(w_{t-1} + v_{t-1})$$

$$(7)\ w_{t-1,t} = (1 - k_t)w_{t-1}$$
$$(8)\ v_t = v_{t-1} + \lambda((m_t - m_{t-1})^2 + w_{t-1} + w_t - 2w_{t-1,t} - v_{t-1})$$

Where the subscript t represents the trial for variables that are updated every trial, k is the Kalman gain, which regulates the update of other variables, w is the estimated (co)variance over m, v is the estimated volatility of the process, ω is akin to observation noise, α is the learning rate for m, m is the estimated outcome (mean) of the process, o is the observation made (reward obtained), s() represents the sigmoid function, and λ is the learning rate for v. m_0, v_0, λ, and ω are parameters given to the VKF. Equations extracted from Piray and Daw, 2020 (Equations 9 to 13). This set of equations still pose that values are updated according to their previous one and a prediction error, so the general principle behind a learning rule is kept.

Equations 9 to 11. Linear ballistic accumulator

$$(9)\ k_s \sim U(0, A)$$
$$(10)\ b = A + B$$
$$(11)\ t_s = \frac{b - k_s}{v} + t_0$$

Where k is the starting amount of evidence for stimulus s, which is sampled from an uniform distribution with limits [0, A] every trial, in which A is the intertrial variability of k, B is the threshold parameter which is added onto A to get the response threshold b, t_s is the reaction time computed for stimulus s, v is the drift rate and t_0 is the non-decision time. Equations deduced from Brown and Heathcote, 2008.

In the VKF models, each stimulus was assigned a VKF to itself, and the m value of the VKF was used as the Q-value for the softmax or multiplied into the drift rate in the LBA to link the learning and decision-making modules of the models.

In this way, the baseline model has in total 2 parameters (α for the simple delta rule and β for the softmax), the stepping stone model has 5 (m_0, v_0, λ, and ω for the VKF and β for the softmax), and the target model has 8 (m0, v0, λ, and ω for the VKF and A, B, v, and t_0 for the LBA).

Parameter Recovery

A parameter recovery analysis was conducted to ascertain whether the VKF is a robust alternative to the simple delta rule, that is, if it recovers well. The parameters for each model where randomly drawn from a uniform distribution with range according to Table 2. A different set of parameters was sampled 50 times, and simulated data of 100 trials was generated for each sampling. For all samplings, only one pair of stimuli was simulated, with reward probabilities of 70% and 30%, which were reversed between the two stimuli at trial 70. Finally, a model was fit through differential evolution with maximum 200 iterations and a parameter space with range equivalent to the initial sampling. Spearman's rank correlation coefficient between the true sampled parameters and the estimated parameters was calculated as a rough indicator of recovery.

Table 2. Uniform distribution range for each parameter for parameter recovery analysis.

Parameter (associated component)	Range
α (simple delta rule)	[0.01, 1]
β (softmax)	[0.01, 5]
m_0 (VKF)	[0.01, 5]
v_0 (VKF)	[0.01, 5]

λ (VKF)	[0.01, 0.99]
ω (VKF)	[0.01, 5]
A (LBA)	[0.01, 5]
B (LBA)	[0.01, 5]
v (LBA)	[0.01, 5]
t_0 (LBA)	[0.2, 1.5]

Descriptive Adequacy

Lastly, the aforementioned models were fit to participant data from Miletić et al., 2021. A full model was fit per participant (N = 46), and blocks with incomplete data (missing RT for any trial) were completely discarded from the dataset. The models were fit with the same differential evolution algorithm, except that the maximum value of parameters which allow a value greater to 1 was bound to 3 instead of 5 (as it is in Table 2). The estimated models were then used to simulate choice (for analysis of the accuracy) and RT (in the case of the model with an LBA) a 1000 times per trial following the order each participant saw the stimuli. The final representative trial, i.e., the response and RT pair selected to represent the typical behavior of the model under the given conditions, was set to the mode of the decision (left or right stimulus) and the mean of the RT. The BIC (Bayesian information criterion; Schwarz, 1978) of each model was calculated to inspect whether the increased complexity of the stepping stone and target models offered a better fit than the mere increase in complexity could explain.

Results

Following the previous section, the results are also shown separated into two subdivisions: parameter recovery and descriptive adequacy. All the data summarized below and the code for the plots can be found in the online repository available at http://www.github.com/mathboger/imcn-uva-internship-2023 .

Parameter Recovery

Figures 2 to 4 show the profile plots of each of the models. Profile plots show the behavior of the likelihood function around the true value of a parameter in generated data. This helps to ascertain the likelihood function is well defined for the type of model in question. On each subplot, all parameters are kept constant but one of interest, and the likelihood of the model as a whole is calculated for that specific change. The value of the parameter of interest is plotted on the horizontal axis, while the likelihood is plotted on the vertical one. A vertical blue line marks the ground truth for that parameter. The data used for the likelihood calculation is an artificially generated one with the constant values of the parameters in the same fashion as described for the parameter recovery in the methods section. As can be seen, the maximum likelihood generally falls around the ground truth, showing that the function used to calculate it is reliable for further analyses.

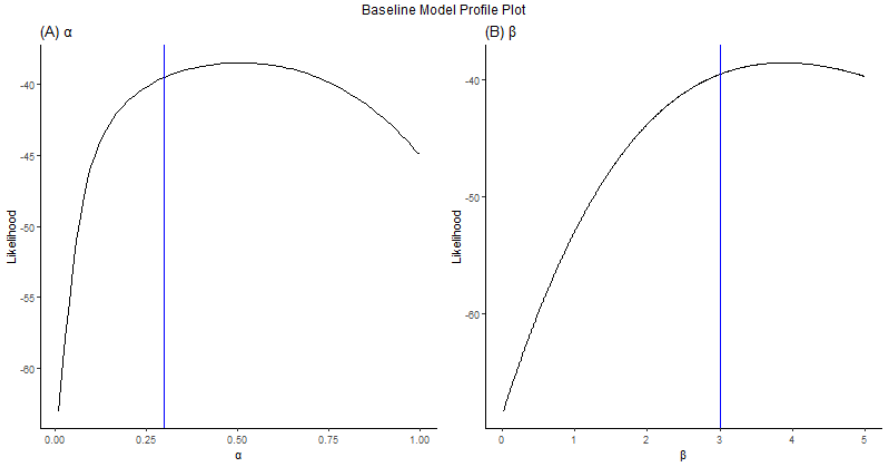

Figure 2. Baseline model profile plot on: (A) α (ground truth of 0.3); (B) β (ground truth of 3.0).

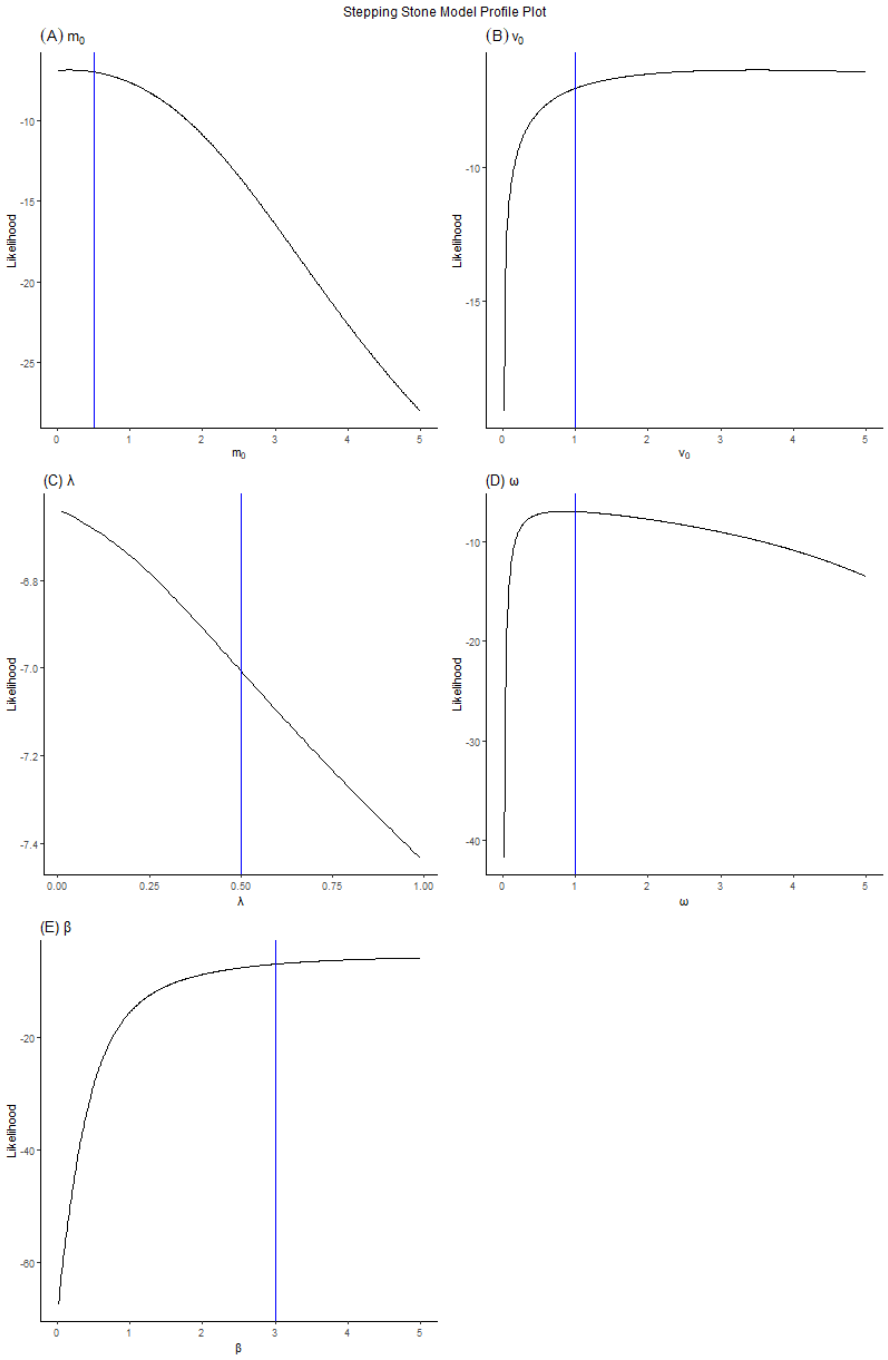

Figure 3. Stepping stone model profile plot on: (A) m_0 (ground truth of 0.5); (B) v_0 (ground truth of 1.0); (C) λ (ground truth of 0.5); (D) ω (ground truth of 1.0); (E) β (ground truth of 3.0).

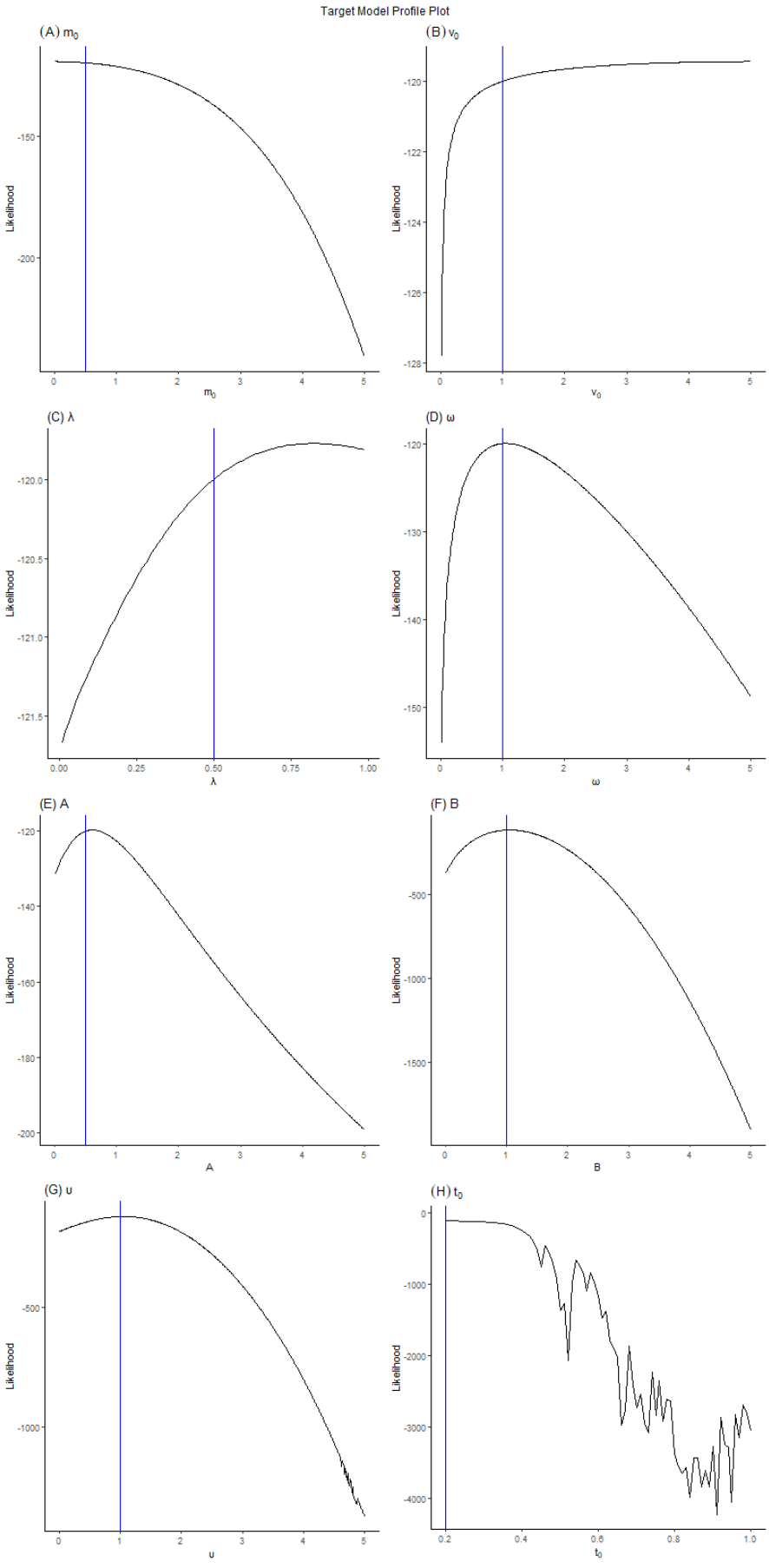
Target Model Profile Plot
(A) m_0
(B) v_0
(C) λ
(D) ω
(E) A
(F) B
(G) υ
(H) t_0
Likelihood
m_0
v_0
λ
ω
A
B
υ
t_0

Figure 4. Target model profile plot on: (A) m_0 (ground truth of 0.5); (B) v_0 (ground truth of 1.0); (C) λ (ground truth of 0.5); (D) ω (ground truth of 1.0); (E) A (ground truth of 0.5); (F) B (ground truth of 1.0); (G) ν (ground truth of 1.0); (H) t_0 (ground truth of 0.2).

Next, Figures 5 to 7 show the result of the parameter recovery analysis for each of the models. For each subplot, a scatterplot of the true value of the parameter in that run (x-axis) is plotted against the estimated value for that same parameter (y-axis). The black line is drawn from a simple linear regression, and if used combined with the blue line, which exemplifies perfect correlation, a visual estimate of how well the parameter is recovered can be achieved. As a rough numeric estimate, the Spearman's correlation coefficient is also shown on the bottom right of each sublot. A recovery of the likelihood is also shown, and as can be observed, the likelihood always recovers well, a sign that the fitting of the models itself is sound.

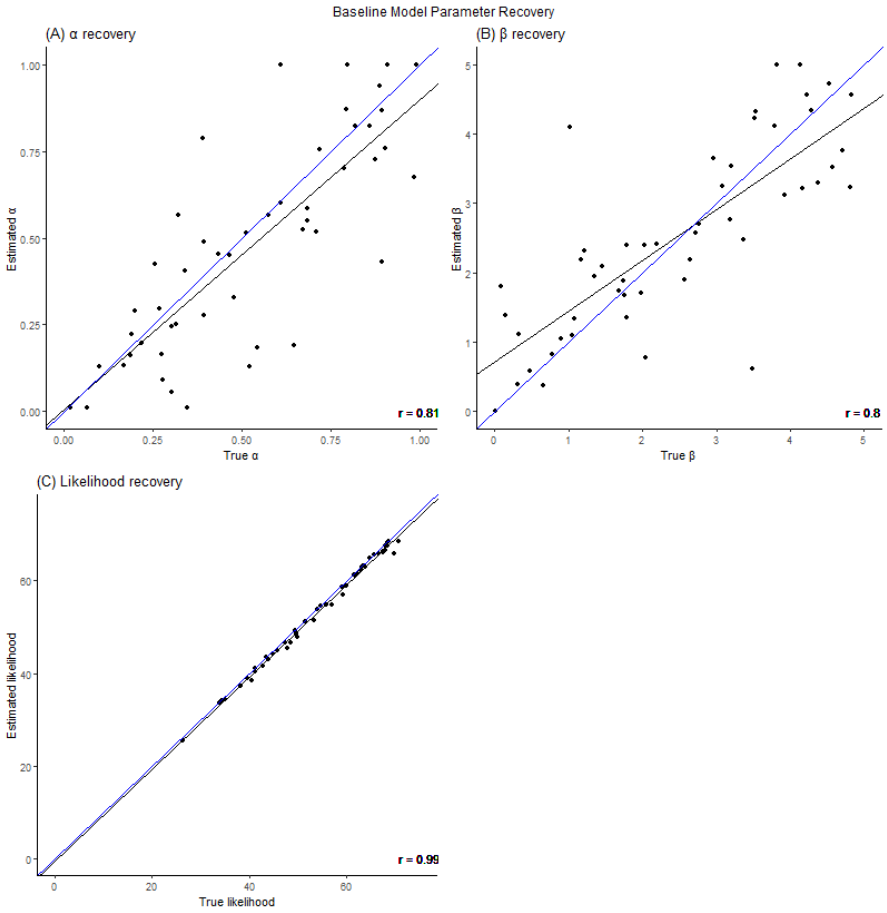

Figure 5. Baseline model parameter recovery plot on: (A) α; (B) β; (C) log-likelihood.

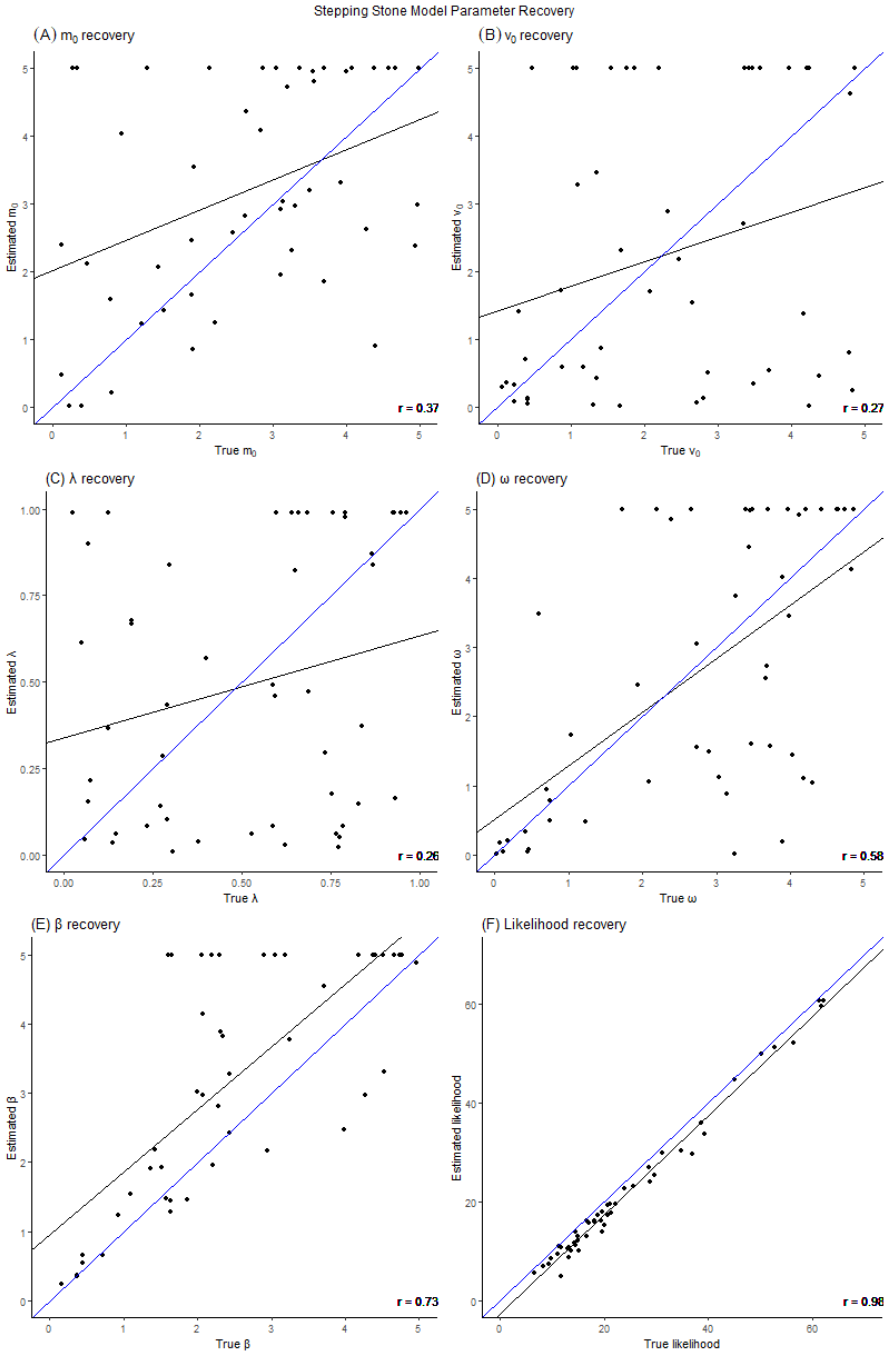

Figure 6. Stepping stone model parameter recovery plot on: (A) m_0; (B) v_0; (C) λ; (D) ω; (E) β; (F) log-likelihood.

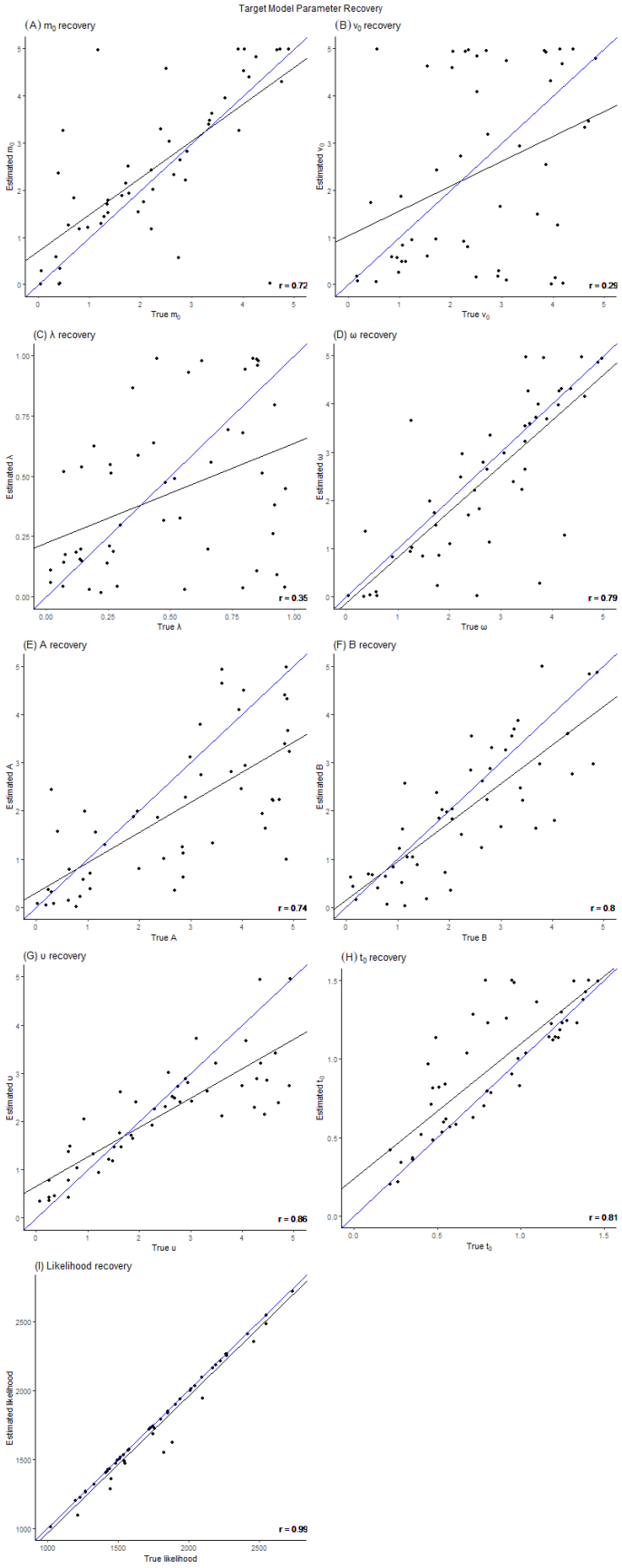
Target Model Parameter Recovery
(A) m_0 recovery
Estimated m_0
True m_0
r = 0.72
(B) v_0 recovery
Estimated v_0
True v_0
r = 0.29
(C) λ recovery
Estimated λ
True λ
r = 0.35
(D) ω recovery
Estimated ω
True ω
r = 0.79
(E) A recovery
Estimated A
True A
r = 0.74
(F) B recovery
Estimated B
True B
r = 0.8
(G) υ recovery
Estimated υ
True υ
r = 0.86
(H) t_0 recovery
Estimated t_0
True t_0
r = 0.81
(I) Likelihood recovery
Estimated likelihood
True likelihood
r = 0.99

Figure 7. Target model parameter recovery plot on: (A) m_0; (B) v_0; (C) λ; (D) ω; (E) A; (F) B; (G) v; (H) t_0; (I) log-likelihood.

Descriptive Adequacy

As for the descriptive adequacy of the models, first Table 3 shows the average fit of each class. Of special interest are the log-likelihood and BIC values, since those can be used to compare the overall fit of each class. Then, Figures 8 to 10 show the results of the data simulation from the estimated parameters against the real participant data aggregated over participants for each block binned in 10 bins of trials of similar size. In those Figures, the horizontal axis represents the bins, while the vertical axis represents the value of the outcome variable of interest. For Figures 8 and 9 the only outcome possible is accuracy, since the softmax function makes no estimation over RTs. This changes for Figure 10, where RTs are plotted separately for correct trials (trials where the participant or the model chose the stimulus with a higher probability of reward) and error trials (trials where the participant or the model chose the stimulus with a lower probability of reward). In all cases, the black line shows the real participant data, with the gray shadowing around it representing the standard error of the data, and the colored line represents the average of the simulated trials (blue for accuracy, green for correct RT and red for error RT). After discarding incomplete data for each block, the final number of participants for block 1 was 44, for block 2 was 42, for block 3 was 40, and for block 4 was 40 for all data shown in the figures.

Table 3. Mean and standard deviation (between parenthesis) across participants of the values of the estimated parameters, log-likelihood and BIC for each model category.

Value / Model	Baseline model	Stepping stone model	Target model
α	0.29 (0.08)	NA	NA
β	2.70 (0.47)	1.00 (0.63)	NA
m_0	NA	0.01 (4e-5)	0.01 (8e-4)
v_0	NA	2.37 (1.14)	2.41 (0.97)
λ	NA	0.49 (0.37)	0.48 (0.38)
ω	NA	1.02 (1.24)	1.21 (1.26)
A	NA	NA	2.25 (0.76)
B	NA	NA	0.50 (0.24)
v	NA	NA	0.83 (0.54)
t_0	NA	NA	0.12 (0.96)
Log-likelihood	-1114.29 (262.03)	-1150.35 (266.44)	-1067.80 (328.91)
BIC	34.91 (1.50)	34.84 (1.55)	35.02 (1.68)

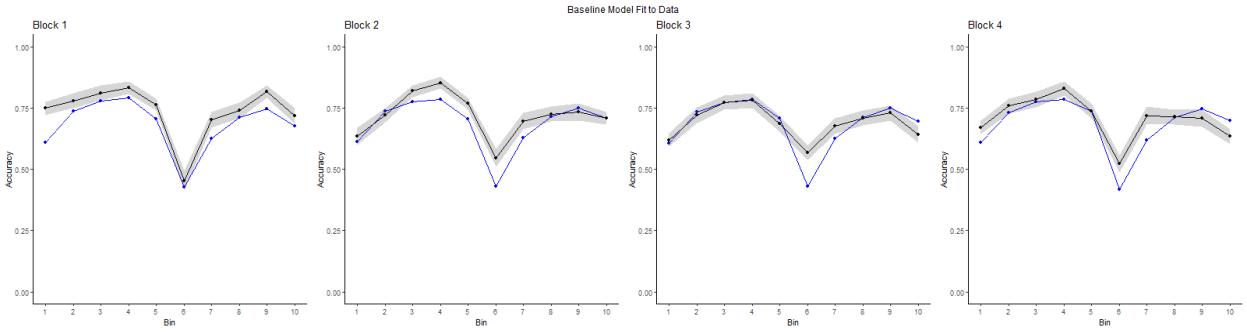

Figure 8. Baseline model descriptive adequacy plot.

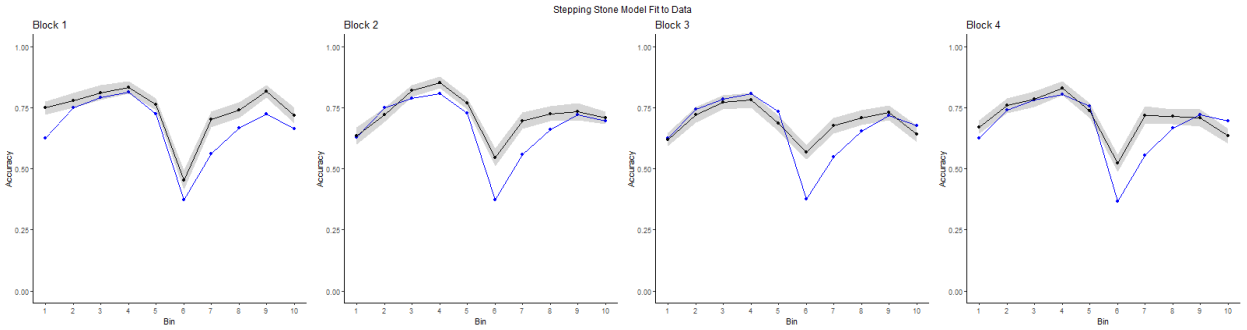

Figure 9. Stepping stone model descriptive adequacy plot.

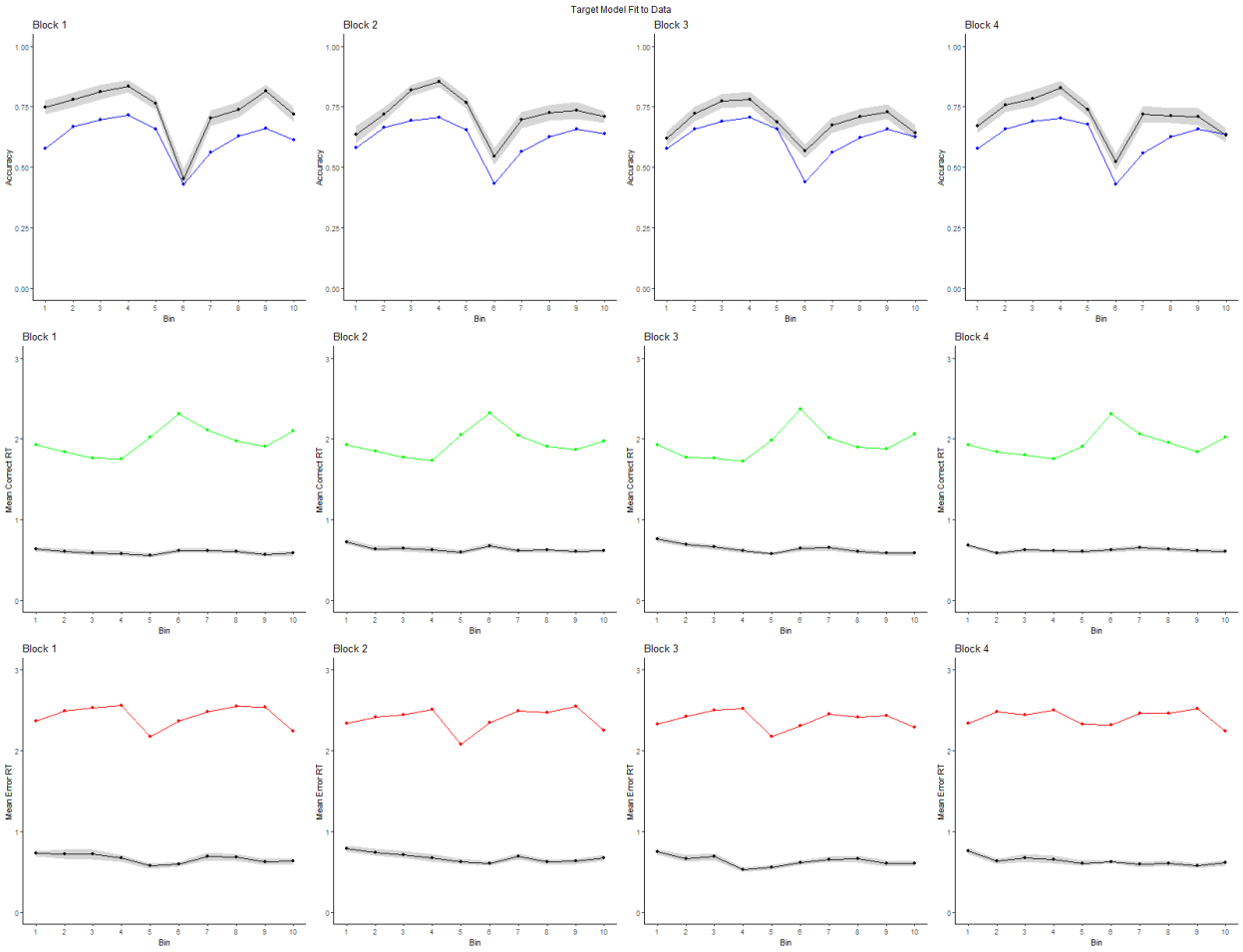

Figure 10. Target model descriptive adequacy plot.

Conclusion

To reiterate the aims of the current study: analyses were conducted to ascertain whether the VKF is a robust method and whether it is a good descriptor of the reversal learning task. To answer the first question, the parameter recovery analysis should be paid attention to. Unfortunately, as can be observed in Figures 6 and 7, the VKF parameters have poor recovery, excepting for ω, which represents the observation noise. The situation is improved for m_0 when the LBA is coupled to the VKF instead of the softmax function, most probably due to the fact that the softmax only looks at the difference between Q-values when calculating the choice probabilities for each stimulus, while for the LBA the absolute value of the Q-value is important since it affects the final drift rate for the trial (Brown & Heathcote, 2008). This makes so that for the stepping stone model to work well, only the difference in m_t for stimuli in a pair is relevant, and since an infinite combination of numbers can yield the same difference, m_0 can vary much more without affecting the suitability of the model. However, since the utmost interest of making use of the VKF is to model the volatility present in the environment (Piray & Daw, 2020), one can argue that the most important parameters to be looked at are v_0 and λ, since they are the initial estimated volatility and the update rate for the estimated volatility. Again, unfortunately, these are the parameters that show overall the worst recovery. This suggests that the VKF as a method on itself is already not very robust prior to its application to real data. It is reasonable to expect of a good mathematical model that it is able to recover to its ground truth quite well when using data generated by itself (Donkin et al., 2011; Forstmann & Wagenmakers, 2015; Palestro et al., 2018), and that was found in this study to not be the case for the VKF. However, there is not only grim news. Looking at the BIC values on Table 3, it is possible to see that even though the VKF adds complexity to the model, it does not worsen the evaluation of fit against complexity. This can be noted in particular through the comparison of the BIC from the baseline and the stepping stone model, where the average BIC is lower for the stepping stone model, with the disclaimer that the values are extremely close to each other, and with a high overlap in their distribution as can be seen from the standard deviation values. This is a relevant finding because it points to the hope that the overarching idea of multi-layered state estimation in the VKF to be sound enough that more robust models based on this principle could still explain the data well without incurring into extreme penalties due to their complexity.

As for the second question, attention must be paid to the descriptive adequacy results. The first issue that comes to eye is the extremely slow RTs displayed by the target model. While the participants exhibit average RTs all lower than the one second mark, the model registers often average RTs higher than two seconds. This basic discrepancy means the fit is too poor for suitable applicability. Most probably, this is an effect of how the RL and the EAM components are linked in the employed model. A simple multiplication of the m_t for the stimulus with a fixed drift rate (v) is how they were linked here, while it is probably wiser to add an intercept or allow trial by trial variation of the drift rate to adjust the overall height of the RT distribution. This intercept coefficient in the drift rate definition is often called an urgency signal, and work on it is available (Cisek et al., 2009; Miletić et al., 2021), with it also being incorporated into the models of Miletić et al., 2021. As for the accuracy, the data attribute that can be compared through all three models, the situation is again not so bright for the target model. All models are able to capture the overall pattern that accuracy dips around the reversal point and then recover to about the previous rate. However, the rate achieved by the target model in the first place is lower than the other models and thus also further away from real participant data. This is also probably due to simplified link between the VKF and the LBA assumed in the current study, since the simple addition of the VKF to the softmax in place of the simple delta rule did not affect the overall similarity in the simulated and real data. A first step in a future study would be to generate simulated data from VKFs associated with EAMs and see how well they are regenerated by an estimated model on top on the ground truth, in a similar fashion to a parameter

recovery analysis. It is expected that such investigation would shed more light on the inner workings of the VKF and how it aligns better with the EAMs already in use for the description of behavior in volatile environments.

On the other hand, the VKF as it is defined by Piray and Daw, 2020, is essentially based on samplings from normal distributions, which are then in the end converted to a binary decision through the application of a sigmoid function. This continuous nature could hinder the applicability of this learning rule to tasks such as the reversal learning task of this study, where the outcome of interest is discrete. A reformulation of the VKF to repurpose it from its inner workings to more discrete decision-making could also be a suitable solution to the issues reported here. However, it must be noted that while making the model a more loyal representation of the task at hand seems alluring, this could also be a problem. If the research on decision-making has its utmost goal on elucidating general processes, then the researchers must be able to find universal patterns that are largely independent of task, and that is precisely one of the reasons that make the VKF stand out. It has a very broad formulation that indicates the possibility of application over many tasks, and once it is better studied, an analysis on how well it applies to the different tasks in the field could elucidate some properties of decision-making under volatility. Even so, this study also calls for some caution when attempting to fit the VKF as it is to many other tasks, since it has shown to not recover so well, and so a reformulation on its precise definitions is probably on demand before such attempts are made. There are also other learning rules that have been proposed to deal with volatility (Gallistel et al., 2014; Mathys et al., 2011), and it would be wise to conduct a similar recovery analysis of those before effort is invested solely into the VKF.

References

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in Changing Conditions: The Urgency-Gating Model. *The Journal of Neuroscience*, *29*(37), 11560–11571. https://doi.org/10.1523/JNEUROSCI.1844-09.2009

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151. https://doi.org/10.1016/j.jmp.2010.10.001

Donkin, C., & Brown, S. D. (2018). Response Times and Decision-Making. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–33). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn509

Forstmann, B. U., & Wagenmakers, E.-J. (Eds.). (2015). *An Introduction to Model-Based Cognitive Neuroscience*. Springer New York. https://doi.org/10.1007/978-1-4939-2236-9

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), Article 1. https://doi.org/10.1146/annurev-psych-122414-033645

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). *The Perception of Probability*.

Mathys, C., Daunizeau, J., Friston, K.J., & Stephan K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00039

Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, *136*, 107261. https://doi.org/10.1016/j.neuropsychologia.2019.107261

Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *ELife*, *10*, e63055. https://doi.org/10.7554/eLife.63055

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20–48. https://doi.org/10.1016/j.jmp.2018.03.003

Piray, P., & Daw, N. D. (2020). A simple model for learning in volatile environments. *PLOS Computational Biology*, *16*(7), Article 7. https://doi.org/10.1371/journal.pcbi.1007963