Task-dependent evaluation of reinforcement-learning models in TAB and DAWH

Aral Cay
PSYC 51: Computational Models of Behavior

November 2025

Abstract

Reinforcement Learning (RL) models are widely used to describe how test subjects learn from different experimental designs and reward outcomes. In this paper, two models are compared on rat behavior in two different tasks. The first one is standard Rescorla-Wagner (RW), and the other is the Stacked Probability (SP) model used in the same learning rule but multiplies the probability of reward on the ignored side of a two-arm setup. Both models have two parameters, alpha (α) and beta (β). I analyzed 383 rat sessions from two-armed bandit (TAB) and dual assignment with hold (DAWH) tasks. RW dominated in TAB (165/215 sessions), whereas SP dominated in DAWH tasks (125/168 sessions). I first checked parameter recovery and found a strong correlation between true and recovered parameters. Model recovery was also performed, and simulations showed that RW and SP were distinguishable with high recoveries. The results supported the idea that rats use different strategies for both tasks.

1 Introduction

This project compared two RL models to understand the decisions rats make in two different tasks (TAB and DAWH). The goal was to determine and test which model works better for each task. In my analysis, I first verified that the parameters can be recovered from the simulated data. Then I tested whether the models can be verified from the simulated data by performing model recovery. The verification was strong, so I fit both models to the real rat data from Shin et al. [1] to compare each model's performance. After fitting, I validated the winning model by comparing the simulated behavior with real data. The two discussed models are explained below:

1. **Rescorla-Wagner (RW) Model:** Standard RW reinforcement learning model. It updates the values according to the prediction errors. It assumes that the animals learn the expected value of each action and choose accordingly.
2. **Stacked Probability (SP) Model:** This is an extension of the standard RW model. It introduces the stacked arming probability to also track the effect of not choosing one arm on the other one. This is assumed to perform better in tasks where the reward structure is based on the subject switching the choice.

Both of these models have two parameters (α - learning rate, β - inverse temperature), which makes AIC and BIC equal for model comparison. These models were fit to two different tasks and simulated [2]. The tasks are ***TAB*** and ***DAWH***.

1. **Two-Armed Bandit (TAB) Task:** There are alternating blocks (each block is approximately 40 trials) with constant reward probabilities on each block. This is a simpler task and would likely favor model free learning.
2. **Dual Assignment with Hold (DAWH) Task:** The reward probability of the unchosen side increases every time it is ignored. This creates an incentive for the rats to switch their choices between arms in the maze.

This makes us ask whether the rats can actually adapt their strategies according to the task. For this, I will test parameter recovery, perform model recovery, fit models to empirical data of both TAB and DAWH, validate the winning models with behavioral measures, and interpret the results.

2 Parameter recovery

I tested whether the parameters can be recovered from the simulated data to justify the model fitting that I performed later. I simulated data with the known parameter values and fit the models to perform parameter recovery. I used 300 trials per session to match the empirical data and for each model, I used a grid of true parameters (Figure 1). For each of the parameter combinations, I simulated 4 datasets, fit the model, and finally compared the recovered parameter with the true parameter. The RW model was tested on TAB, and SP was tested on DAWH task data.

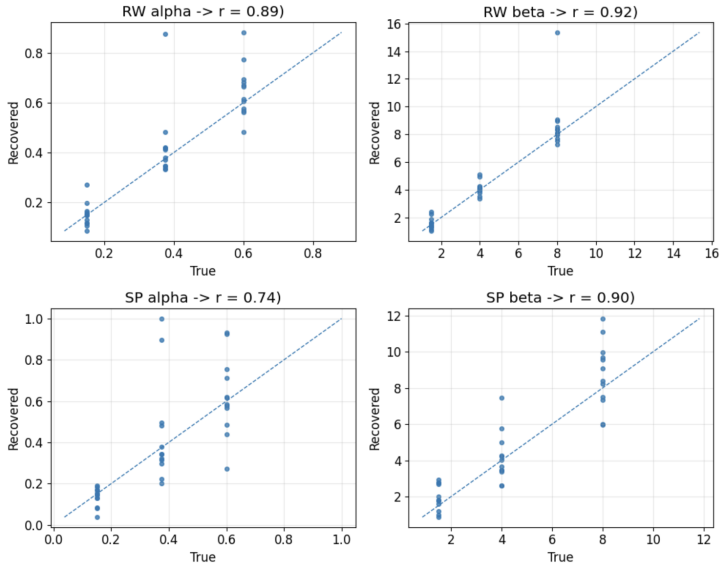


Figure 1: Parameter recovery for RW and SP. Each point is a simulated session (300 trials)

Figure 1 shows the recovered values plotted against the true values. The points cluster around the identity line, especially stronger for β. So, this means that the Pearson correlations between the true and recovered parameters were strong:

- RW α: $r = 0.89$
- RW β: $r = 0.92$
- SP α: $r = 0.74$
- SP β: $r = 0.90$

The correlations show that the parameters can be recovered and identified with the amount of data available.

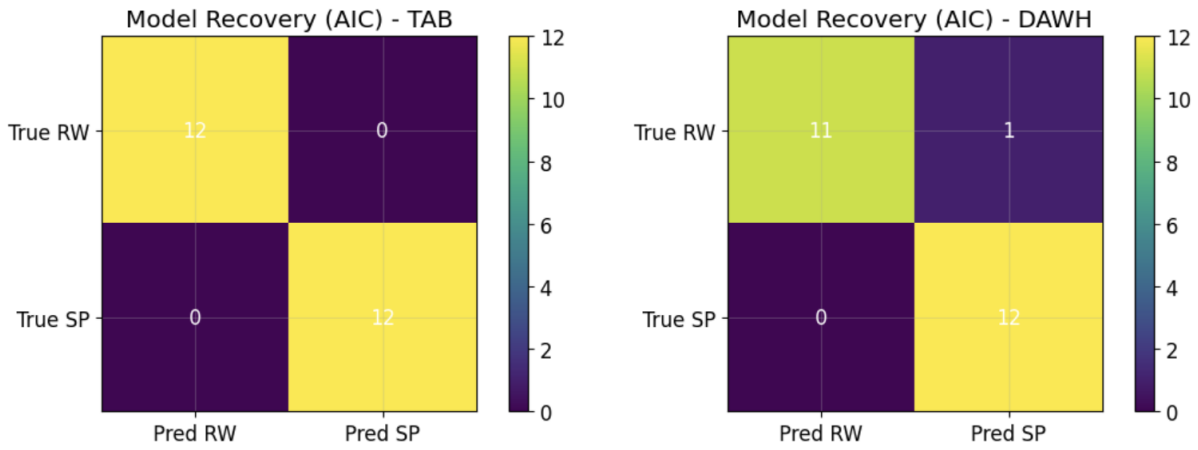
3 Model recovery

I tested if I can identify the data generating model correctly when fitting to simulated data. This would validate that the models are distinguishable and model comparison can be performed. I generated 24 datasets for each task (12 from RW and 12 from SP) and fit both models to each of the datasets (Figure 2) to use AIC and BIC to select the winning model. Since there were two parameters in both models, they gave the same result, so only the AIC results will be presented. I created confusion matrices and observed that model recovery was highly accurate for both of the tasks.

- **TAB:** 100% accuracy (12/12 correct RW, 12/12 correct SP)
- **DAWH:** 95% accuracy (11/12 correct RW, 12/12 correct SP)

Figure 2 shows the confusion matrices for TAB and DAWH. The overall recovery was strong for both, which means that the model comparison on real data would be reliable. So if either SP or RW were the true data generating process, then AIC would do a good job comparing and pick the correct model most of the time.

The confusion matrix of the task TAB was perfectly diagonal, so AIC predicted all tasks correctly. However, there was a small asymmetry in the DAWH task. All of the datasets generated by the SP model were correctly predicted but 1 out of 12 RW generated sets were classified as SP. This was probably due to SP having access to both learned values and arm probabilities. With some of the parameter values, it is expected to look similar to simpler learning behaviors.



(a) TAB task **(b)** DAWH task

Figure 2: Model recovery confusion matrices (AIC)

4 Parameter fitting to empirical data with model comparison

I fitted both of the RL models (RW and SP) to real rat data from various research that explored the TAB and DAWH sessions [3–8]. The dataset includes a total of 383 sessions with 215 TAB and 168 DAWH task sessions. I fit the models independently at session level and compared the models using AIC. For each session, I fit both of the models with maximum likelihood estimation (MLE). I performed AIC for each model and also computed delta logarithmic likelihood (SP - RW). Figure 3 shows the distributions of the best fits of α and β.

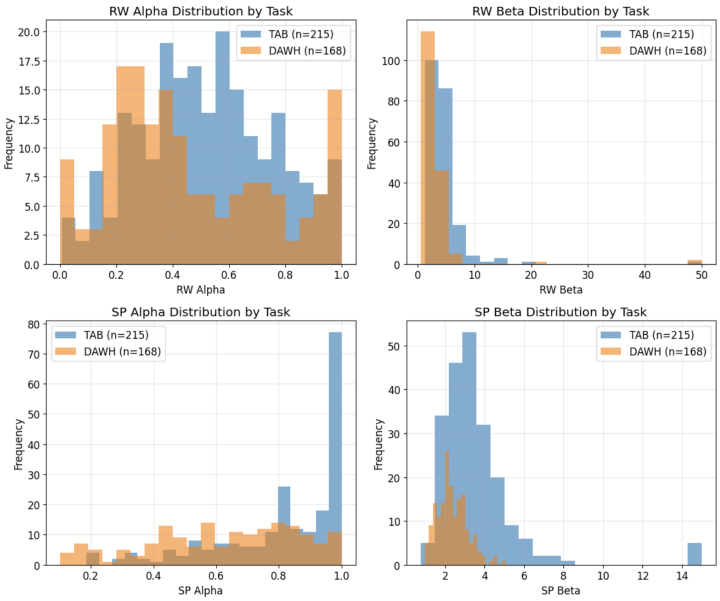


Figure 3: Distributions of best-fit α and β by task for RW and SP models

Model comparison

For each session I also computed AIC for both models and chose the winner model by the lower value. Looking at Table 1, we can see that RW clearly dominates TAB task sessions and SP clearly dominates DAWH tasks. The chi square test on this table gives $\chi^2(1) = 97.38$ and $p < 10^{-16}$ which means model preference was strongly dependent on the task.

Table 1: Contingency table of winning models by task (AIC).

	RW wins	SP wins
TAB task (215 sessions)	165	50
DAWH task (168 sessions)	43	125

In all of the real data sessions the mean delta log likelihood was slightly negative, showing that RW won more sessions overall. In TAB delta log likelihood is strongly negative on average, and in DAWH, it was strongly positive. To compare the delta log likelihoods, I used Mann Whitney test and compared between the tasks and I got $p < 10^{-4}$, and the standardized effect size (Cohen's d) was 1.32. All these results showed that RW model was a better explanation for TAB and SP was better suited to DAWH.

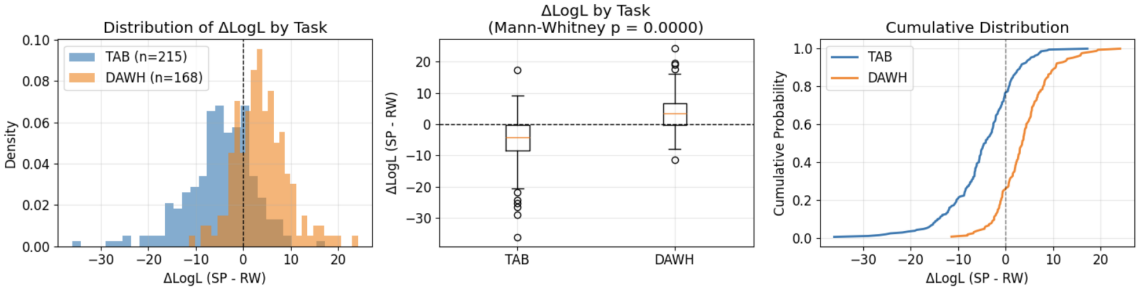


Figure 4: Task-level comparison of model fits using $\Delta LL = LL_{SP} - LL_{RW}$.

5 Model validation

After getting the results, I compared the simulated behavior with real data and validated the winning model (Figure 5). For every session, I took the model that won AIC and used the best fit parameters, simulated new choice and reward of the same length, then computed model independent behavioral measures. Then these measures were correlated between sessions to see how well the models reproduced the behaviors.

- Win–stay rate: $r = 0.90$
- Lose–shift rate: $r = 0.76$
- Reward rate: $r = 0.96$
- Switch rate: $r = 0.87$
- Mean run length: $r = 0.79$

For each of the empirical session, I picked the AIC winner model and its best fit parameters. Then, I simulated 10 new sessions to compute the behavioral measures on the simulated data. I averaged across the simulations and plotted the correlations between real and simulated values across sessions.

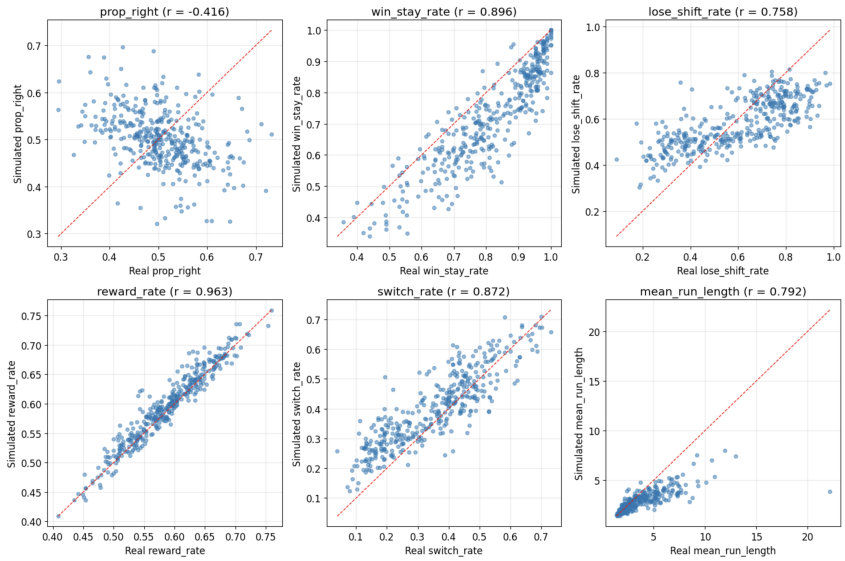


Figure 5: Model validation using model-independent summary statistics. Each point is a session.

In Figure 5, we can see that points cluster around the identity line. The only exception was the proportion of the right arm choice, the correlation was r = -0.42. The tasks are symmetric

with respect to right and left, so the absolute proportion of the right arm choices just weakly constrained by the fit. The sessions with left or right bias can look equally likely under the models. All of the other behavioral aspects were captured smoothly by the models.

5.1 Run-Length Analysis

I performed model-independent analysis of choice probability as a function of run length to see whether the models capture behavioral aspects emphasized by Shin et al. [1]. I focused on the last 20 trials from each block to see how the probability of choosing the left arm changes with the number of consecutive same action choices. I computed and plotted the relationship between real data and the models (Figure 6).

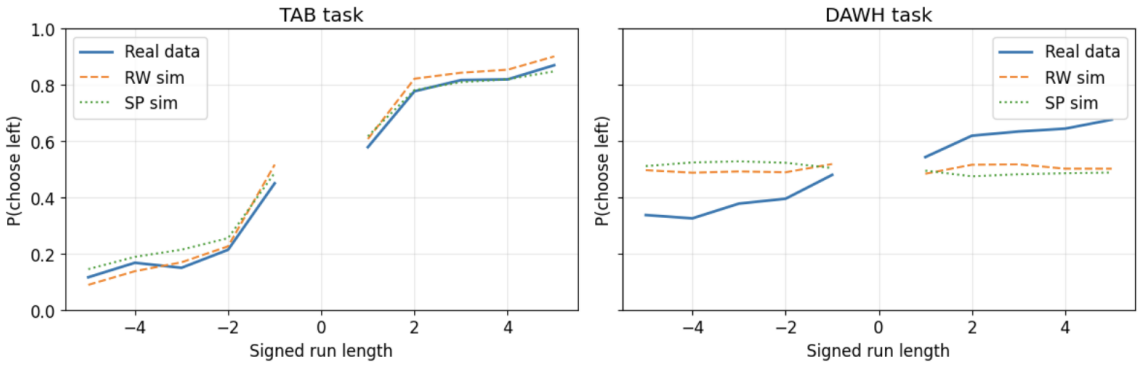


Figure 6: P(choose left) vs signed run length for TAB and DAWH comparing real sessions

For TAB task, the curve shows a strong dependence on the run length, and both RW and SP show this pattern. In DAWH, the real data shows a positive relationship between the run length and choice probability of the better side but both models produced curves that are flat with weak dependence on the run length. This means that the SP wins in DAWH in the likelihood analysis, but models could not capture the specific way the rats adjust the choices in long runs in DAWH.

6 Questions

6.1 Do different tasks favor different cognitive models?

Does DAWH task engage more sophisticated run length tracking and actually favor SP, while the simpler TAB task can be relied on RW? If the cognitive strategies the rats chose were task adaptive, we should see a divergence that shows DAWH dependence on SP and TAB dependence on RW. Looking at Table 1, we can see that RW wins 165/215 TAB sessions, and DAWH sessions favored SP 125/168 of the trials. The chi square test gave me $\chi^2(1) = 97.38$ and $p < 10^{-16}$, which shows that the model preference depends on the task. This means that the model comparison results supported the task dependence discussed in [2].

6.2 How do RL parameters differ between tasks?

Both of the models had two parameters α and β and across both models, β was generally larger in TAB than DAWH. This is probably because the TAB environment is stable within the blocks so the logical choice would be to stay at the better arm for the session. DAWH changes the

reward probabilities of the ignored side, so a behavior that tries to explore would be more logical for the task.

The learning rate α was mostly higher in SP. When SP is used for TAB task, it just makes up for the extra structure of stacking by learning fast, so it performs an RW like strategy. For DAWH, both models showed a bit lower learning rates. So, this means the learning rules were different in the two tasks.

7 Discussion

I compared two RL models discussed by Huh et al. [2] on two decision making tasks TAB and DAWH. Both models showed good parameter and model recovery. Then, I fit the models to 383 empirical sessions and compared them using AIC and delta log likelihood. RW performed considerably better in the TAB sessions with around 76.7% and SP performed better in DAWH tasks with 74.4%. Model validation also showed that the winner models were capturing the model independent behaviors like reward rate, win stay / lose shift, switching behavior, and run length.

7.1 Confusing results

I also got some unexpected results, which may need more investigation in future work:

1. **SP and run length effect:** SP wins for the DAWH task strongly, but the run length curves I got were flat compared to the real data. This may mean that the SP model may have some missing mechanisms like choice biases for long runs.
2. **Probability of choosing the right arm:** There was a negative correlation for proportion of the right arm choices (Figure 5). In the tasks and setup, left and right identification is not necessary, so this would not cause a problem in understanding the tasks and models.
3. **Parameter Boundaries:** The parameter estimates clustered around the boundaries. This could mean either the parameters at boundaries were actually genuine or the parameter bounds were a bit restrictive.

7.2 Final Thoughts

If I had more time and resources to work on this project, I would first want to explore and test some variants or alternatives of the SP model that may take the learning rule more as a function of run length. I would then test how the conclusions may change when I change the parameter bounds and use other model comparison metrics.

With this project, I tried to perform systematic computational modeling to test, validate, and compare the results from Huh et al. and Shin et al. [1, 2]. Working on a computational modeling pipeline to validate the results made the ideas and discussions more concrete for me to understand. With this work, I saw that the RL models can reveal the task dependent strategies and differences in choice and behavior of rats.

References

[1] Shin, E.J. *et al.* (2021) 'Robust and distributed neural representation of action values', *eLife*, 10. doi:10.7554/elife.53045.

[2] Huh, N. *et al.* (2009) 'Model-based reinforcement learning under concurrent schedules of reinforcement in rodents', *Learning & Memory*, 16(5), pp. 315–323. doi:10.1101/lm.1295509.

[3] Kim, H. *et al.* (2009) 'Role of striatum in updating values of chosen actions', *The Journal of Neuroscience*, 29(47), pp. 14701–14712. doi:10.1523/jneurosci.2728-09.2009.

[4] Lee, S.-H. *et al.* (2017) 'Neural signals related to outcome evaluation are stronger in CA1 than CA3', *Frontiers in Neural Circuits*, 11. doi:10.3389/fncir.2017.00040.

[5] Sul, J.H. *et al.* (2010) 'Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making', *Neuron*, 66(3), pp. 449–460. doi:10.1016/j.neuron.2010.03.033.

[6] Sul, J.H. *et al.* (2011) 'Role of rodent secondary motor cortex in value-based action selection', *Nature Neuroscience*, 14(9), pp. 1202–1208. doi:10.1038/nn.2881.

[7] Kim, H., Lee, D. and Jung, M.W. (2013) 'Signals for previous goal choice persist in the dorsomedial, but not dorsolateral striatum of rats', *The Journal of Neuroscience*, 33(1), pp. 52–63. doi:10.1523/jneurosci.2422-12.2013.

[8] Lee, H. *et al.* (2012) 'Hippocampal neural correlates for values of experienced events', *The Journal of Neuroscience*, 32(43), pp. 15053–15065. doi:10.1523/jneurosci.2806-12.2012.