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. 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. These RL models captured some features of rat behavior but there was no strong evidence showing that different tasks favor different cognitive strategies based on likelihood based model comparison.

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, 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.

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, Pearson correlations between the true and recovered parameters were strong:

- RW α: r = 0.89
- RW β: r = 0.93
- SP α: r = 0.89
- SP β: r = 0.95

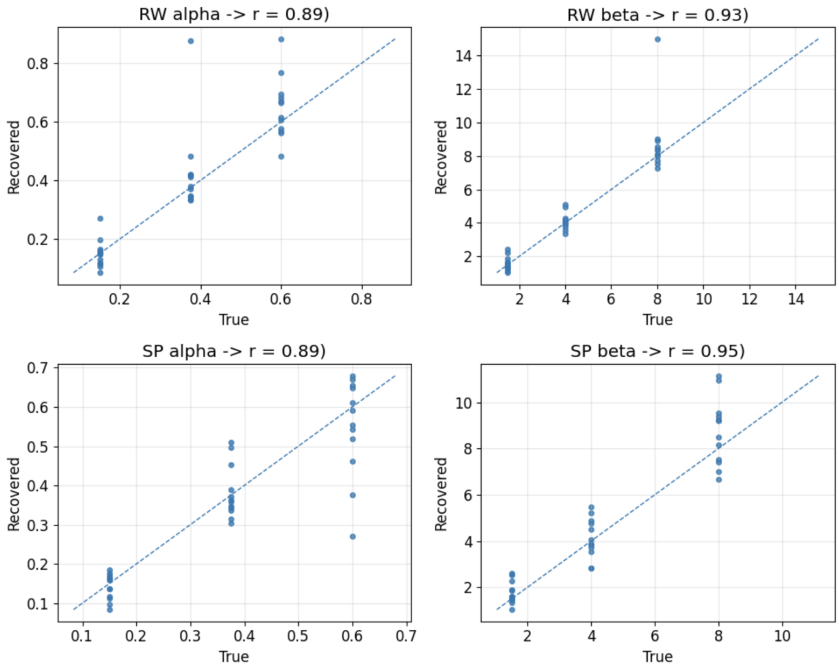
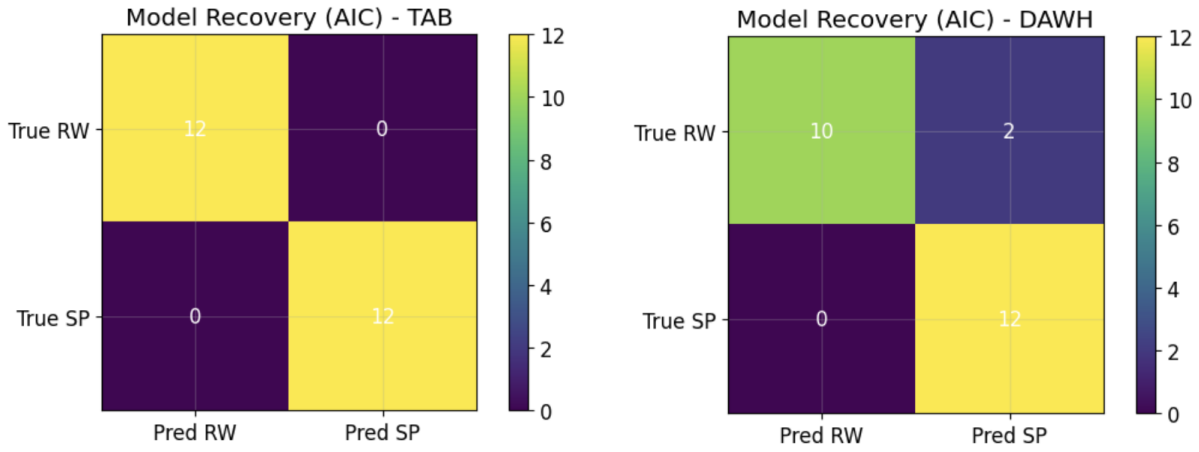


Figure 1: Parameter recovery for the RW and SP models. True vs recovered parameter values for simulated sessions (300 trials)

The correlations show that the parameters can be recovered with the amount of data available. It is important to note that I have observed some clusters hitting the optimization bounds (α near 0 or 1 and β near 0.5 or 15). This means that the bounds may be a bit restrictive, but this does not invalidate the results. Parameter recovery scatter plots (Figure 1) show the true vs. recovered parameter values for all four combinations. Points cluster around the diagonal line which confirms that the recovery was good with some outliers.



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

Figure 2: Model recovery confusion matrices (AIC)

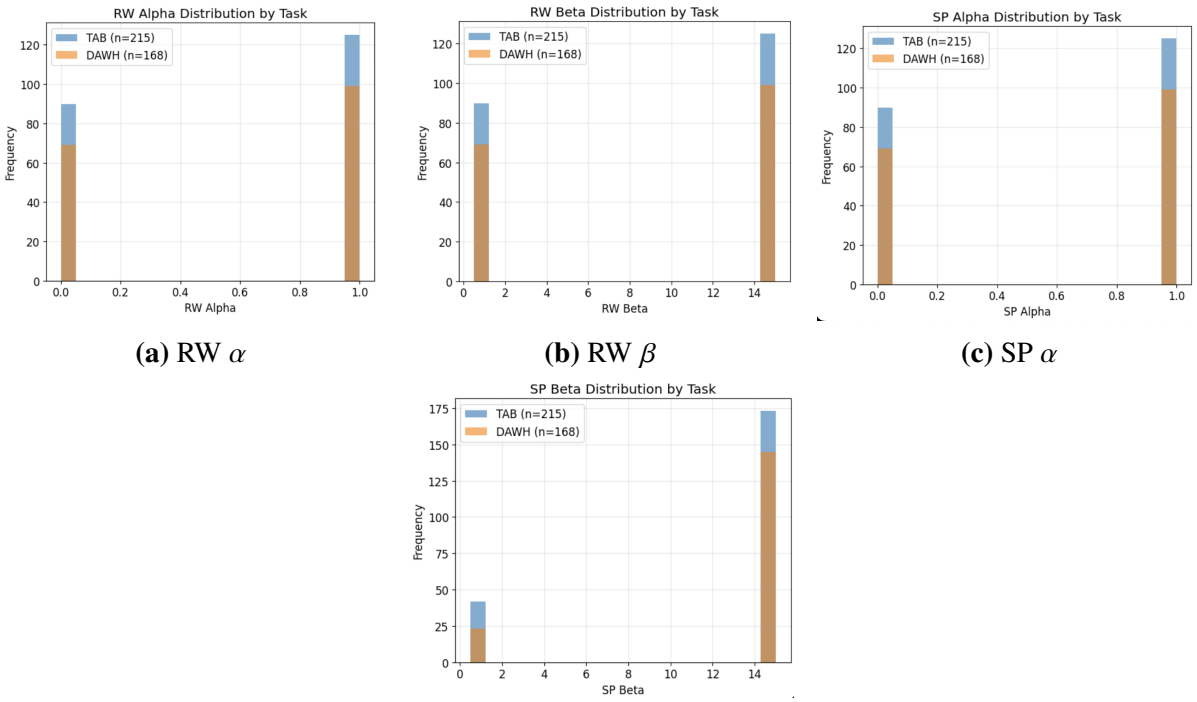
3 Model recovery

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:** 91% accuracy (10/12 correct RW, 12/12 correct SP)

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 2 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) RW α **(b)** RW β **(c)** SP α

(d) SP β

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

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 compute delta logarithmic likelihood (SP - RW). There were again some clusters around the boundaries for both model parameters (Figure 3) which suggests that the parameters could be a bit restrictive. However, with further tests I performed using different bounds for α and β, there was evidence that for many of the sessions the parameter values were genuinely around the bounds. The contingency table shows that TAB favors RW and DAWH favors SP (Table 1).

Table 1: Contingency Table - AIC

	RW wins	SP wins
TAB task:	114	101
DAWH task:	81	87

5 Model Validation

After getting the results, I compared the simulated behavior with real data to validate the winning model (Figure 4). I tested whether the model can reproduce the important aspects of the real behavior. For each of the sessions, I took the winning model based on AIC and the best fit

parameters. For a randomly selected 10 sessions, I computed the six behavioral measures for real and simulated data. The correlations between real and simulated data were not confidently significant, especially for some measures, but moderate:

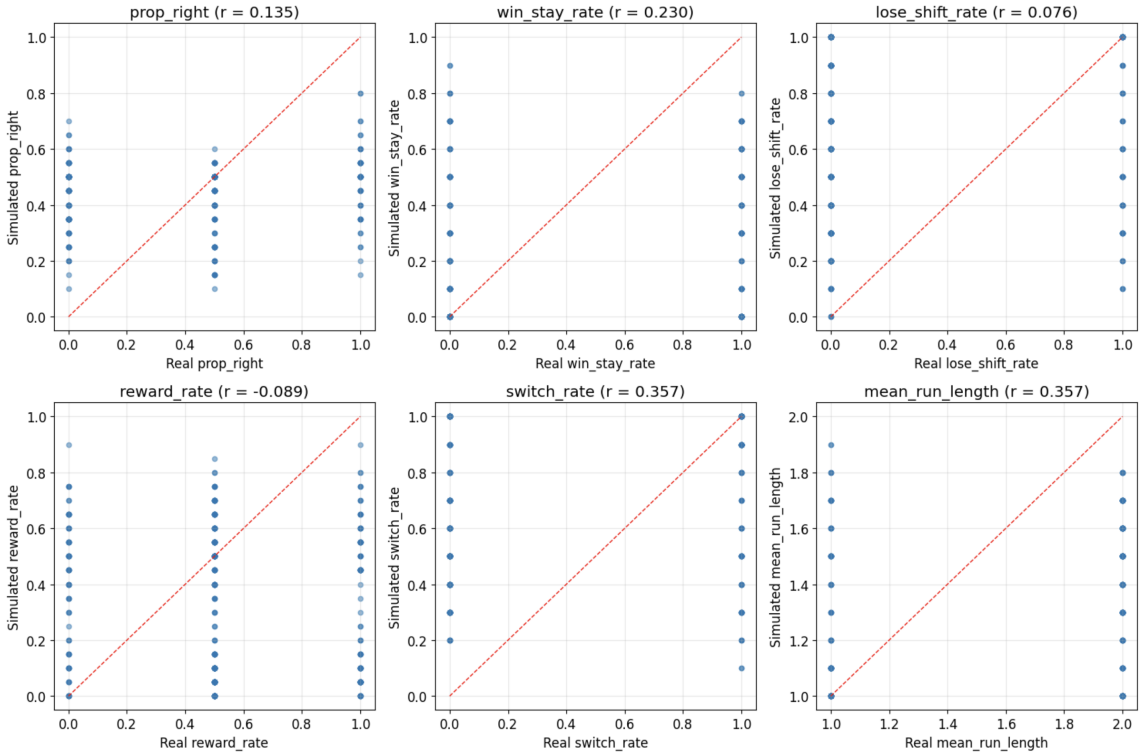


Figure 4: Model validation using model-independent summary statistics.

- Choice Proportions (r= 0.135)
- Win-Stay Rate (r= 0.230)
- Lose-Shift Rate (r= 0.076)
- Reward Rate (r= -0.089)
- Switch Rate (r= 0.357)
- Mean Run Length (r= 0.357)

The strongest correlations were in switch rate and mean run length, so the models captured some behavioral patterns, but they did not reproduce all aspects perfectly. Weak validation in correlations does not necessarily mean that computational models are invalid. However, this shows that the models may have some limitations in characterizing the rat decision behavior. The results show that the RL models were good at capturing the frequency of the switch and how long they keep choosing a certain arm, but the models missed some detailed win-stay lose-switch (WSLS) behavior and reward rate. This was expected because both models are simple two parameter models and there are probably additional factors affecting the behavior.

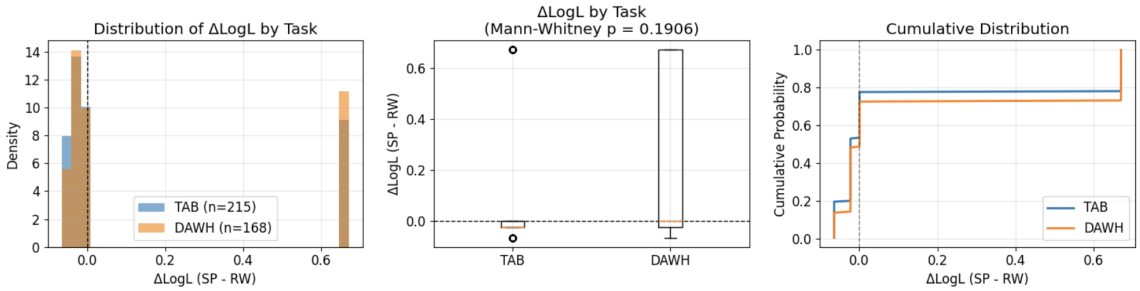


Figure 5: Task-level comparison of model fits using $\Delta LL = LL_{SP} - LL_{RW}$.

5.1 Model-Independent 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).

For TAB task, real data shows good relationship with run length and choice probability. Both of the models captured the pattern well, but SP had a steeper decline, and RW was closer. For DAWH, the real behavior showed a negative slope, and SP simulations also showed a decrease in the slope. This indicates that the even likelihood based model comparison gives a result close to 50-50, the SP model captures the behavioral aspect that separates DAWH from TAB.

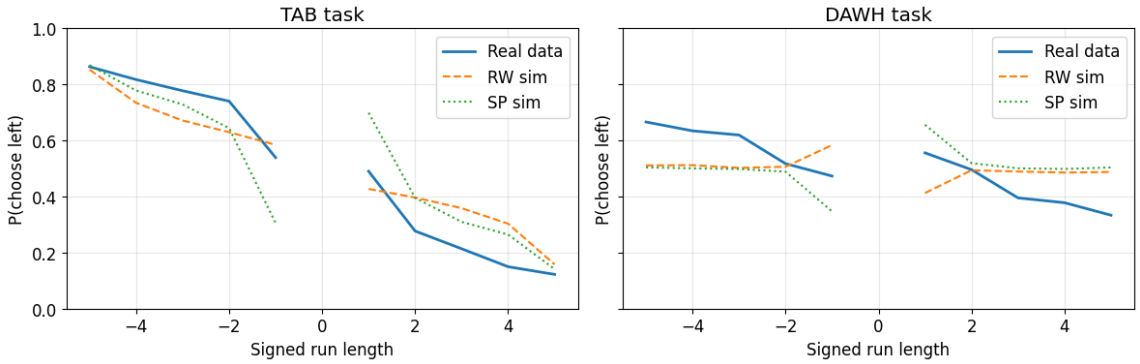


Figure 6: P(choose left) vs signed run length for TAB and DAWH comparing real sessions

6 Questions

6.1 Do different tasks favor different cognitive models?

Does DAWH task engage more sophisticated run length tracking and actually favor SP and 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 Figure 5, this hypothesis was tested but neither of the models identified a clear advantage: RW won 53% of the TAB sessions and SP won 52% of DAWH sessions, and by chi-square test, I got $\chi^2 = 0.691$ and p = 0.406 that showed the difference was not as significant as expected. However, the relationship between run length and choice probability shows that in

DAWH task, only SP model reproduces the negative slope observed in real behavior as shown in Figure 6. This may show the limitation of relying only on the likelihood for model selection.

This close split could indicate that the individual rats used different strategies or the models captured different aspects of the same mechanism. This suggests that the correct decision-making may be captured somewhere between both models.

6.2 How do RL parameters differ between tasks?

Do rats adjust their learning rate or decision making sensitivity in different reward tasks? I used Mann-Whitney U tests to compare the parameters α and β between TAB and DAWH.

Table 2: Parameter Comparison

Parameter values by task		
TAB task (n=215 sessions)		
RW Alpha:	M=0.581,	SD=0.494
RW Beta:	M=8.930,	SD=7.170
SP Alpha:	M=0.581,	SD=0.494
SP Beta:	M=12.167,	SD=5.762
DAWH task (n=168 sessions)		
RW Alpha:	M=0.589,	SD=0.493
RW Beta:	M=9.045,	SD=7.155
SP Alpha:	M=0.589,	SD=0.493
SP Beta:	M=13.015,	SD=4.999

The α values are nearly identical (Figure 3) and β values show a really minor difference. Also, when we look at the standard deviations, we can confirm that the parameter distributions will be similar. This supports the claim that these two tasks did not require different learning speeds.

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. In the real data, RW performed slightly better in TAB task sessions and SP performed slightly better in DAWH task sessions. The models captured some of the behavioral patterns, but they were not able to reproduce perfect behavioral results.

7.1 Confusing Results

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

1. **Model-Independent Run Length:** RW and SP produced similar likelihoods but different behavioral signatures, which made me question if referring only to likelihood is enough for model selection.

2. **Model Balance:** The models went through the expected direction for tasks, but the sessions were split to nearly 50-50 for SP and RW for both tasks in the likelihood analysis. I was expecting a greater distinction that shows that RW would favor TAB and SP would favor DAWH. The models may be capturing different aspects of the cognitive tasks that rats performed. This may also suggest that we may need other approaches to understand the decision making behind these tasks.
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.
4. **Model Validation:** The correlations were moderate and lower than the ideal value. The models can get a good fit even with moderate validation, which I assumed was due to likelihood based fitting capturing different aspects of the behavior.

7.2 Final Thoughts

If I had more time and resources to work on this project, I would first address boundary clustering by testing with different bounds and using different optimization methods. I would also like to test additional models not discussed by Huh et al. [2] to measure and capture task behaviors. Together with alternative models, I would then like to test a hybrid model of RW and SP.

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]. The models captured some aspects of cognitive behavior but the results may need more investigation as discussed above.

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.