Supplementary Information for

Beyond model-based and model-free learning: irrational learning of successor features in addictive behaviour

Table of Content

Supplementary Table S1: the participants' demographical characteristics

Table 1. Participant characteristics

	HC (n = 43)	SUD (n = 83)	Test
Age[a]	34.622 (8.547)	32.524(8.377)	$t = 1.332$, $p = 0.186$
Sex: female/male	0/45	0/ 83	
Education (years)[b]	11.860(2.356)	7.917(3.275)	$t = 7.479$, $p < 0.001$
Nonverbal IQ: WASI-matrix reasoning[c]	17.115(5.989)	9.537(4.889)	$t = 5.410$, $p < 0.001$
Depressive symptoms: SDS[d]	37.342 (5.601)	45.013(6.330)	$t = -6.612$, $p < 0.001$
Anxious symptoms: SAS[d]	32.342(5.828)	40.182(7.869)	$t = -6.017$, $p < 0.001$
Impulsive symptoms: BIS[d]			
planning impulsiveness	22.477(6.741)	32.000(10.310)	$t = -5.935$, $p < 0.001$
motor impulsiveness	22.079(6.895)	23.260(8.491)	$t = -0.7989$, $p = 0.427$
cognitive impulsiveness	23.263(5.254)	32.844(9.274)	$t = -7.057$, $p < 0.001$
Mental disorders: DSM-5 level 1[d]			
depression	2.105(1.429)	2.039(1.618)	$t = 0.224$, $p = 0.823$
anger	0.711(0.694)	1.052(0.902)	$t = -2.240$, $p = 0.027$
mania	2.316(1.317)	1.974(1.469)	$t = 1.259$, $p = 0.212$
anxiety	2.211(1.961)	2.844(2.201)	$t = -1.564$, $p = 0.122$
somatic distress	1.474(1.606)	2.260(1.860)	$t = -2.340$, $p = 0.022$
suicidal ideation	0.289(0.611)	0.584(0.817)	$t = -2.169$, $p = 0.033$
psychosis	0.658(1.097)	1.299(1.540)	$t = -2.564$, $p = 0.012$
sleep disturbance	1.132(1.119)	1.039(0.993)	$t = 0.433$, $p = 0.666$
memory	0.553(0.760)	0.909(0.846)	$t = -2.277$, $p = 0.025$
repetitive thoughts and behaviors	1.474(1.720)	1.961(1.705)	$t = -1.433$, $p = 0.156$
dissociation	0.579(0.889)	0.753(0.845)	$t = -1.005$, $p = 0.318$
personality functioning	1.632(1.847)	1.571(1.743)	$t = 0.168$, $p = 0.867$
substance use	1.132(1.580)	2.456(2.075)	$t = -3.794$, $p < 0.001$
Drug craving: DDQ[d]			
drug desire and intention	-	16.636(8.340)	
drug negative reinforcement	-	9.818(5.358)	
drug control	-	5.078(2.742)	
Abstinence duration (day)[d]	-	205.649(334.277)	
Treatment duration (month)[d]	-	9.445(6.182)	
Pre-treatment drug use (g/day)[d]	-	2.216(3.071)	
Comorbidity (yes/no)[l]	0/45	8/69	

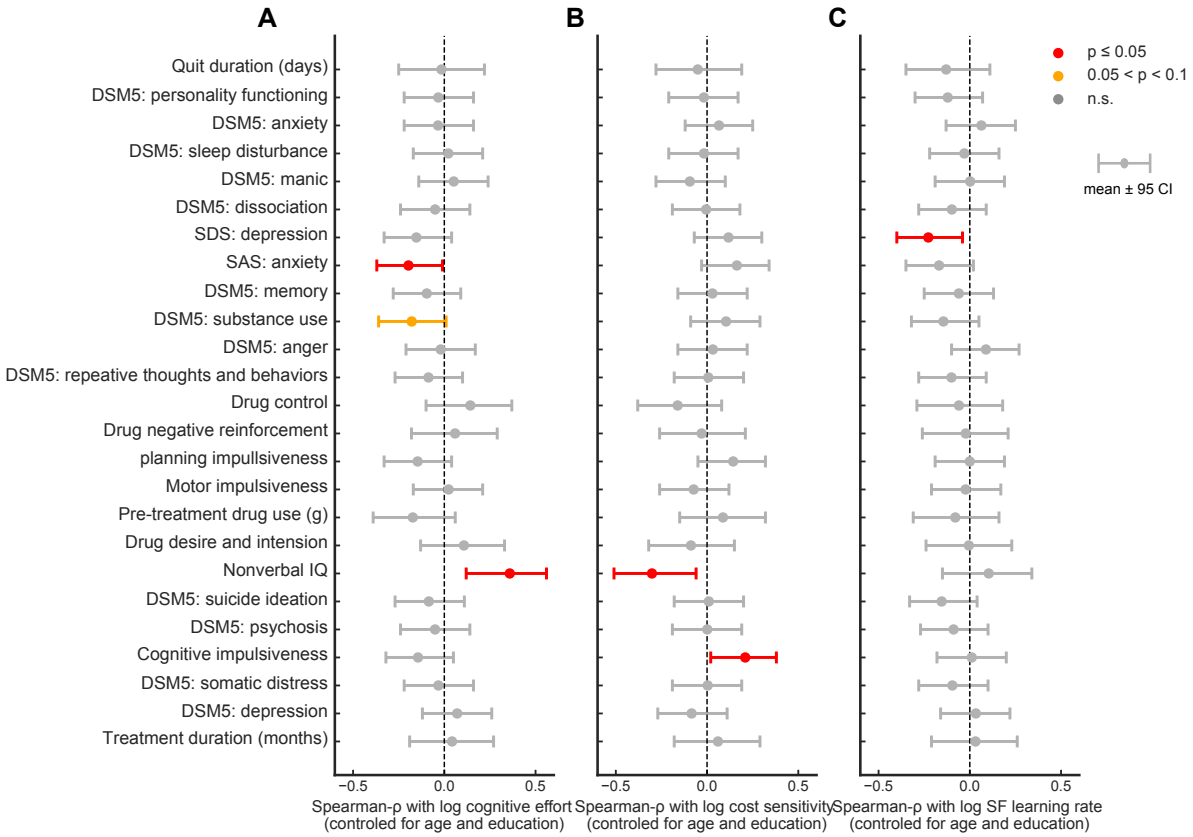
Note: Italicized items represent sub-dimensions of the corresponding scale.

[a] Age data are missing for n = 1 participants with SUD.

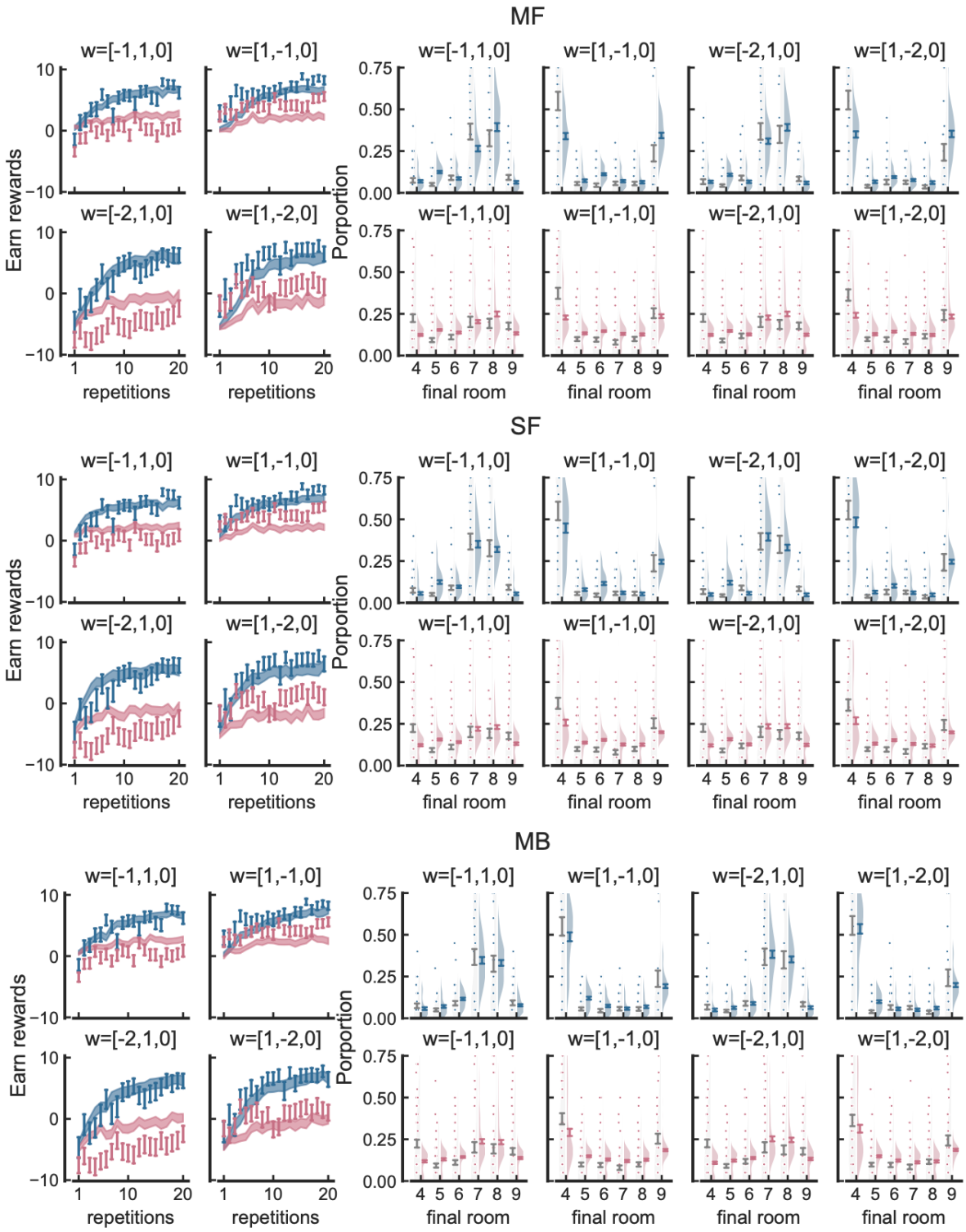
[b] Education data are missing for n = 2 participants with HC, n = 11 participants with SUD.

[c] Nonverbal IQ data are missing for n = 19 participants with HC, n = 42 participants with SUD.

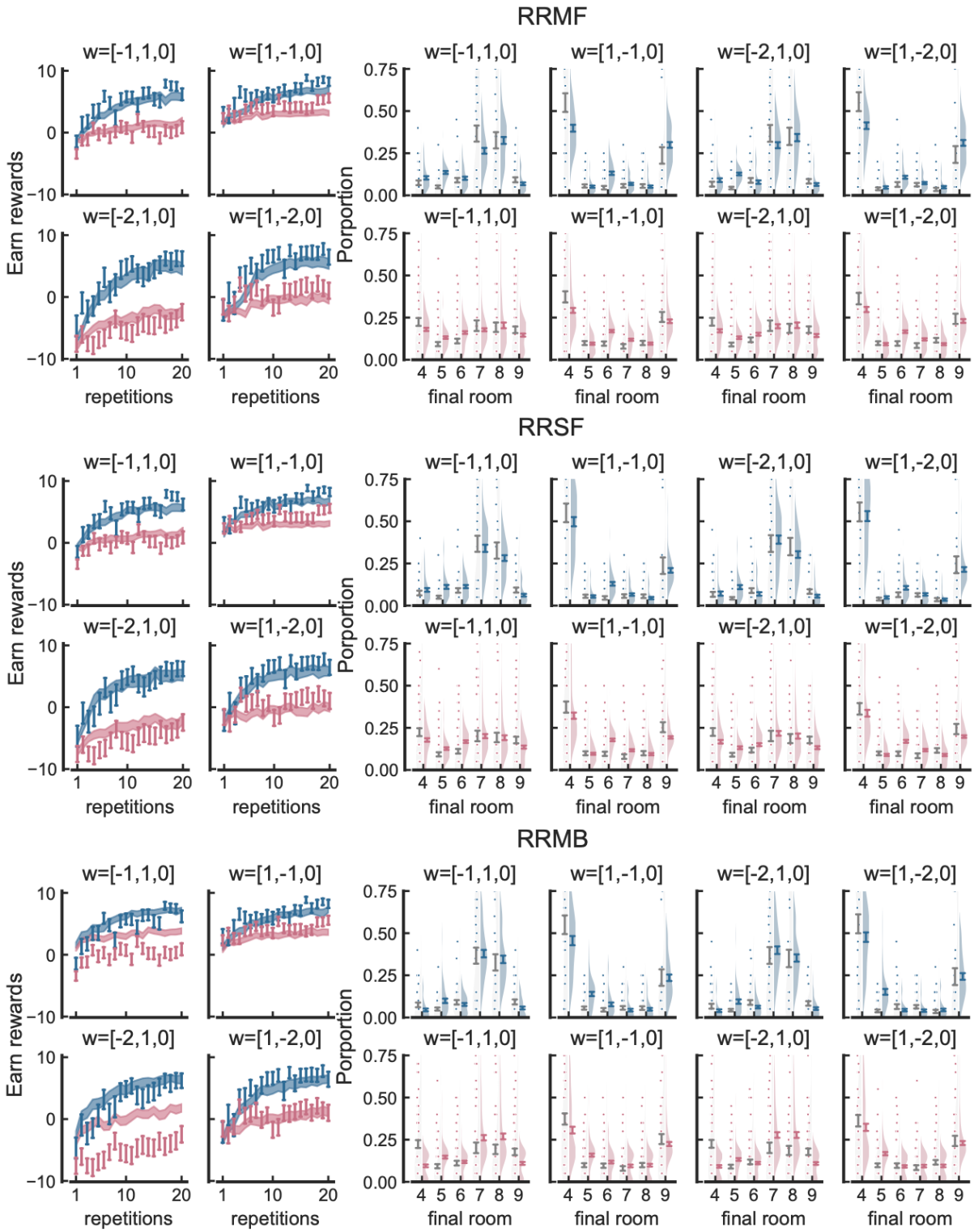
[d] data are missing for n = 7 participants with HC, n = 6 participants with SUD.



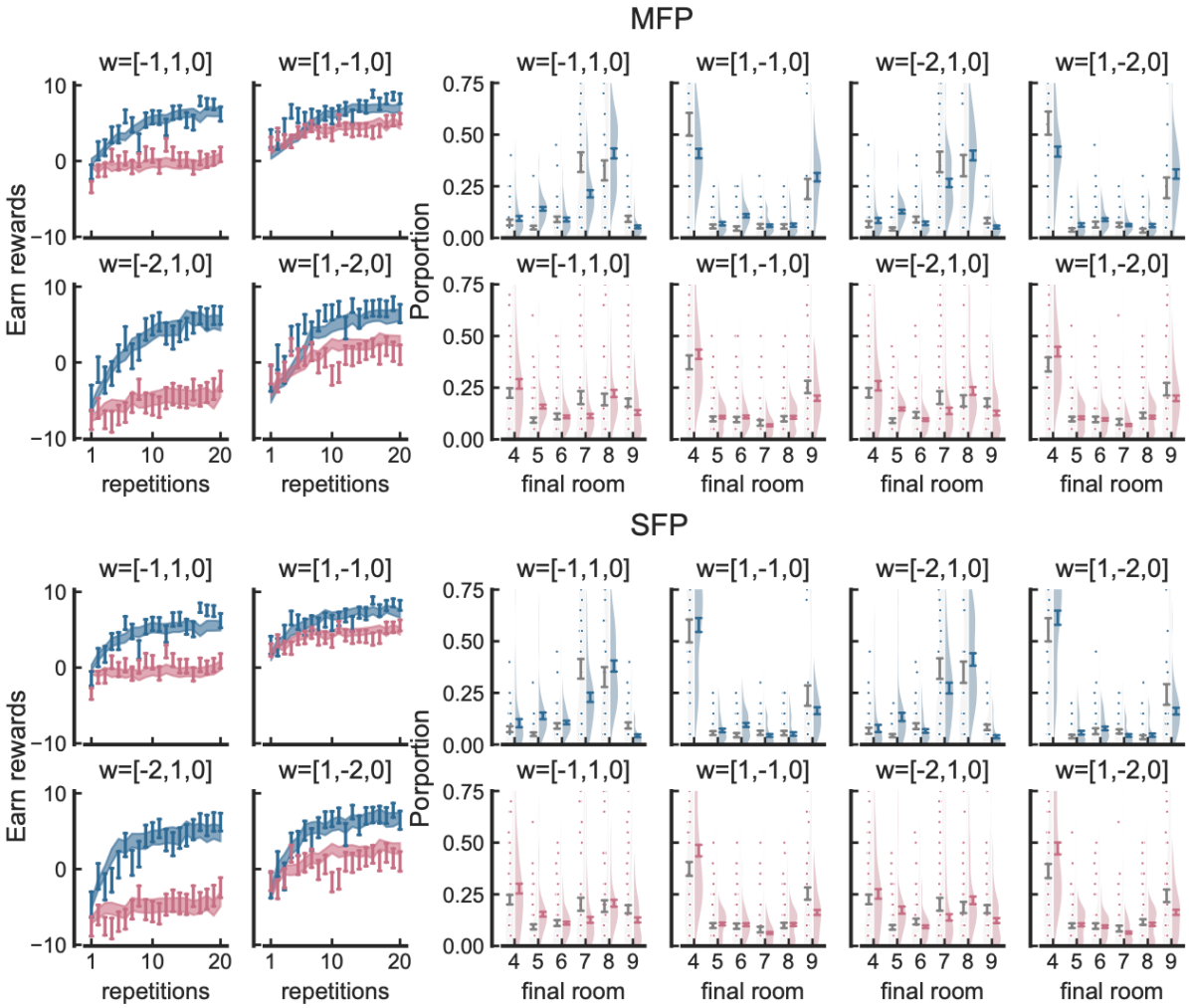
Supplementary Figure S1. the correlation between policy complexity and participants' characteristics. Partial Spearman correlation analysis between participants' **A)** (log) cognitive effort, **B)** (log) cost sensitivity and **C)** (log) successor feature learning rate and their characteristics, controlling for age and education level. The estimated correlation coefficient and its 95% confidence interval are reported here. The significant correlations are in red.



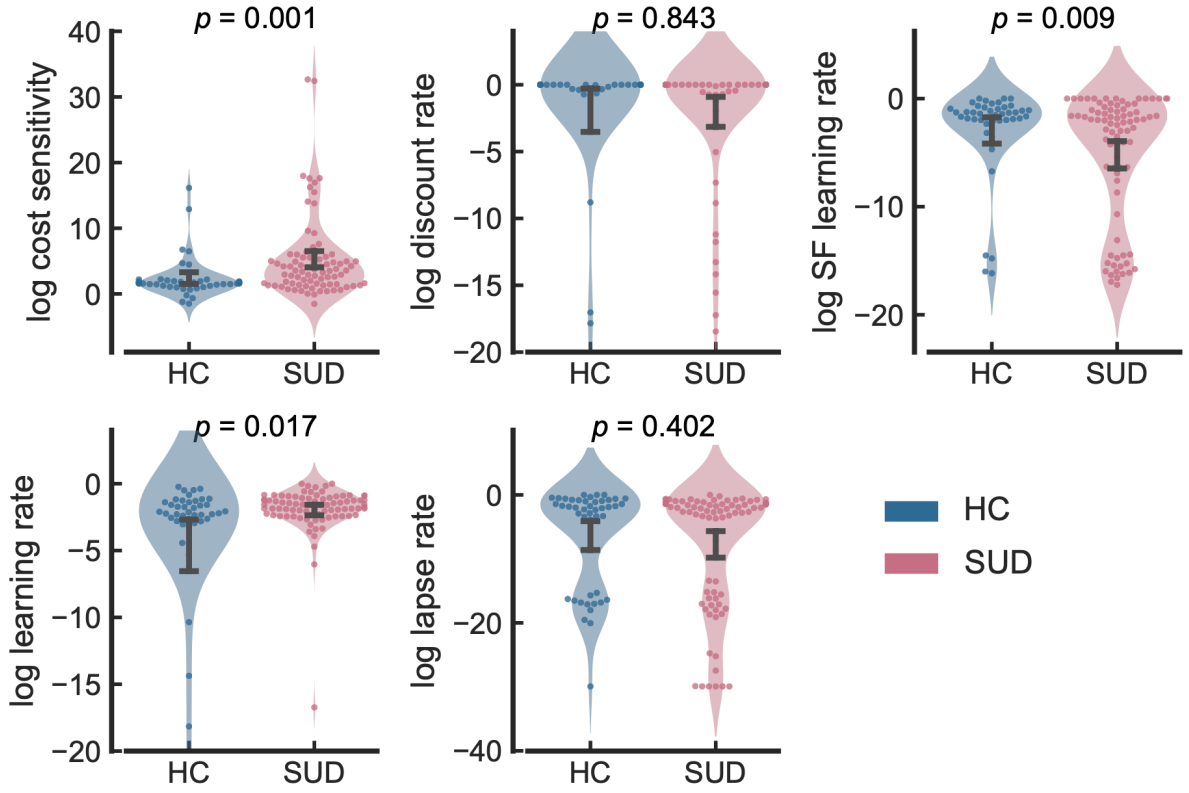
Supplementary Figure S2. Behaviors of Classical RL models.



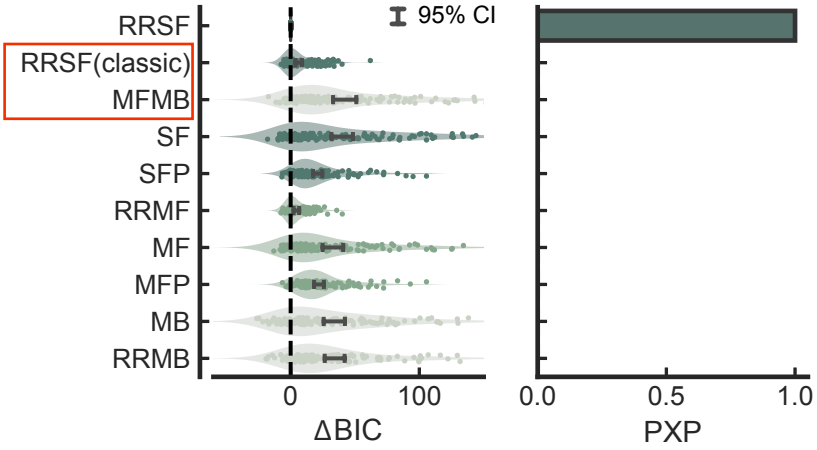
Supplementary Figure S3. Behaviors of Resource-Rational models.



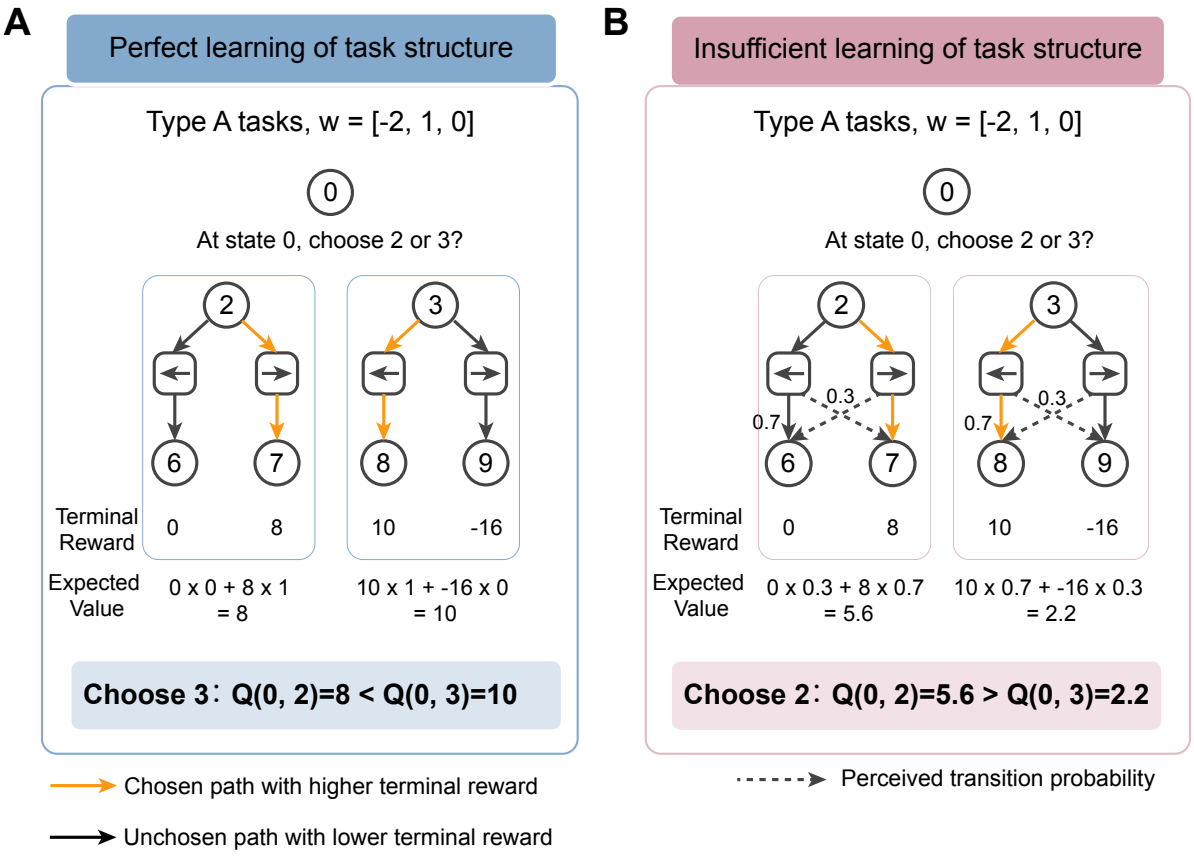
Supplementary Figure S4. Behaviors of Classical RL model with perseveration.



Supplementary Figure S5: RRSF parameter differences between HC and SUD groups.



Supplementary Figure S6: Model Comparison with two potential alternatives.



Supplementary Figure S7. Choose the 2nd best room under insufficient learning of task structure. Illustration of why individuals may prefer the second-best room under insufficient learning of task structure in Type A tasks.

A. Ideally, individuals should learn that choosing "right" in state s = 2 deterministically leads to s = 7, and choosing "left" in s = 3 deterministically leads to s = 8. In this case, the value of s = 2 would be lower than that of s = 3, and individuals should prefer s = 3 at the first step, ultimately reaching s = 8.

B. However, under insufficient learning, individuals may fail to learn the deterministic transitions and instead maintain a stochastic belief—e.g., that choosing "right" in s = 2 leads to s = 7 with 0.7 probability and to s = 6 with 0.3 probability. Because s = 9 incurs a large negative cost in Type A tasks, the expected value of s = 3 (which may lead to s = 9) becomes lower than that of s = 2. As a result, individuals tend to choose s = 2 at the first step, leading them to s = 7, the second-best room.

Supplementary Note 1: Analytical solution for value iteration in resource-rational model-based reinforcement learning

The central step of value iteration in the resource-rational framework is to find the policy π to optimize the following formula,

$$V(g,s) = \max_{\pi} \sum_{a} \pi(a|g,s) \sum_{s'} T(s'|s,a) \left[r(s';g) - \lambda \log \frac{\pi(a|g,s)}{\pi_0(a|s)} + \gamma V(g,s') \right]$$

Following ref.[1], this objective can be written as the constrained optimization problem,

$$\max_{\pi} \sum_{a} \pi(a|g,s) \sum_{s'} T(s'|s,a) \left[r(s';g) - \lambda \log \frac{\pi(a|g,s)}{\pi_0(a|s)} + \gamma V(g,s') \right], \text{s. t.} \sum_{a} \pi(a|g,s) = 1$$

This yields the following Lagrangian

$$J_\eta(\pi) = \sum_{a} \pi(a|g,s) \sum_{s'} T(s'|s,a) \left[r(s';g) - \lambda \log \frac{\pi(a|g,s)}{\pi_0(a|s)} + \gamma V(g,s') \right] + \eta \left(\sum_{a} \pi(a|g,s) - 1 \right)$$

Note that $\pi_0(a|s) = \sum_g p(g)\pi(a|g,s)$, we optimize the Lagrangian via an alternating scheme[2]:

$$\begin{cases} \pi(\cdot\,|g,s) = \operatorname{argmax}_{\pi} J_\eta(\pi) \;\; (\text{with } \pi_0 \text{ fixed}) \\ \pi_0(a|s) \;\; = \sum_g p(g)\pi(a|g,s) \end{cases}$$

The policy step is solved by setting the gradient to zero, $\partial J_\eta(\pi)/\partial\pi(a|g,s) = 0$, together with the normalization constraint $\sum_a \pi(a|g,s) = 1$.

$$\frac{\partial J_\eta(\pi)}{\partial \pi(a|g,s)} = \frac{\partial \sum_a \pi(a|g,s) \sum_{s'} T(s'|s,a)[r(s';g) + \gamma V(g,s')]}{\partial \pi(a|g,s)}$$

$$- \frac{\partial \lambda \sum_a \pi(a|g,s)[\log \pi(a|g,s) - \log \pi_0(a|s)]}{\partial \pi(a|g,s)} + \frac{\partial \eta \sum_a \pi(a|g,s)}{\partial \pi(a|g,s)}$$

$$\frac{\partial J_\eta(\pi)}{\partial \pi(a|g,s)} = \sum_{s'} T(s'|s,a)[r(s';g) + \gamma V(g,s')] - \lambda(\log \pi(a|g,s) + 1 - \log \pi_0(a|s)) + \eta$$

Let $\frac{\partial J_\eta(\pi)}{\partial \pi(a|g,s)} = 0$, we obtain

$$\log \pi(a|g,s) = \frac{1}{\lambda} \sum_{s'} T(s'|s,a)[r(s';g) + \gamma V(g,s')] + \log \pi_0(a|s) - 1 + \frac{\eta}{\lambda}$$

Because $\frac{\eta}{\lambda} - 1$ is not a function of action a and thus can be absorbed into the multiplier η, introducing $\tilde{\eta} = \frac{\eta}{\lambda} - 1$. We then obtain the solution

$$\pi(a|g,s) = \frac{1}{Z} \exp\left(\frac{1}{\lambda} \sum_{s'} T(s'|s,a)[r(s';g) + \gamma V(g,s')] + \log \pi_0(a|s) \right)$$

where the normalization term $Z = \frac{1}{\exp(\tilde{\eta})}$. This formular is equivalent to,

$$\pi(a|g,s) \propto \pi_0(a|s) \exp\left(\frac{1}{\lambda}\sum_{s'} T(s'|s,a)[r(s';g) + \gamma V(g,s')]\right)$$

Since the task-state-action value can be defined as,

$$Q(g,s,a) = \sum_{s'} T(s'|s,a)[r(s';g) + \gamma V(g,s')]$$

The policy can be rewritten as,

$$\pi(a|g,s) \propto \pi_0(a|s) \exp\left(\frac{1}{\lambda} Q(g,s,a)\right)$$

This formula applies to all resource-rational models, including RRMF and RRSF. Once we compute the Q-value function, we can use the formula to obtain the resource-rational policy.

References

1 Tishby, N. & Polani, D. Information theory of decisions and actions. in *Perception-action cycle: Models, architectures, and hardware* 601-636 (Springer, 2010).

2 Boyd, S. P. & Vandenberghe, L. *Convex optimization*. (Cambridge university press, 2004).