STATISTICAL LEARNING AND DATA SCIENCE (SLDS)

Statistical Modeling for Reward Learning and Decision-making for Human Behavioral Tasks

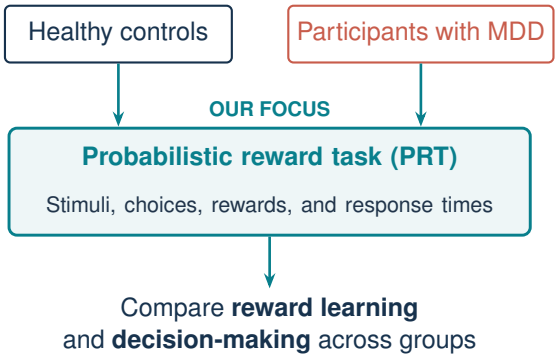
Xingche Guo

Department of Statistics, University of Connecticut

New York / November 1, 2026

EMBARC: a setting for behavioral phenotyping

EMBARC is a **large-scale, multisite randomized clinical trial** investigating biomarkers of antidepressant response in major depressive disorder (MDD).



Complementary data

- Demographic and clinical data
- Task and resting-state EEG/fMRI

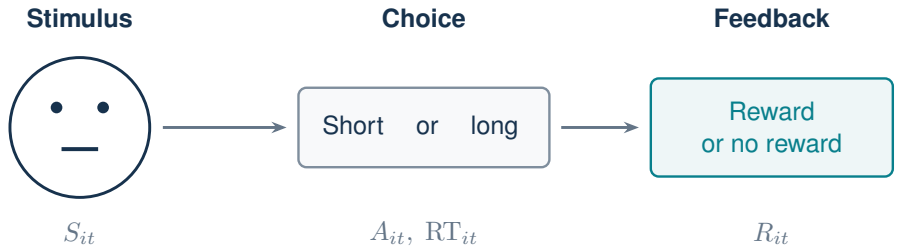
Within the MDD group

Sertraline versus placebo (randomized treatment).

Trivedi et al. (2016). EMBARC: rationale and design.

The probabilistic reward task

Identify a briefly presented mouth as **short** or **long**.



Participant
Identify the stimulus and earn rewards.

Analyst
Infer how the participant learns and decides.

Pizzagalli, Jahn, and O'Shea (2005).

Subtle stimuli, asymmetric rewards

Two nearly identical faces, presented briefly and equally often.



Short mouth: rich

75%

chance of reward if correct

Only mouth length changes



Long mouth: lean

30%

chance of reward if correct

The participant is **not told the reward rule**.

Incorrect choices receive no reward; correct choices are **not always rewarded**.

Illustrative stimuli and reward rule. EMBARC allocates rewarded correct trials in a 3:1 rich-to-lean ratio. Pizzagalli et al. (2005); Guo et al. (2024).

From reward behavior to mechanisms

In the PRT, we observe **choices**, **response times**, and **rewards**.

We want to understand how participants **learn and decide**.

Scientific question	Modeling focus
How quickly do choices adapt to feedback?	Reward learning
How strongly do rewards guide choices?	Reward sensitivity
Does task engagement change over time?	Changes in engagement

Beyond choices

Response times and other tasks provide additional evidence to help distinguish these mechanisms.

- **Goal:** connect behavioral differences in MDD to underlying learning and decision processes.

Trial history is the data

For subject i, observe a sequence of **stimuli**, **choices**, **response times**, and **rewards**.

Trial	$t-2$	$t-1$	t	$t+1$	$t+2$
Stimulus S	Long	Short	Short	Long	Short
Choice A	Long	Short	Short	Short	Short
Response time RT (ms)	620	540	510	690	560
Reward R	0	1	1	0	0

$$\mathcal{H}_{it} = \{(S_{ij}, A_{ij}, \mathrm{RT}_{ij}, R_{ij}) : j < t\}, \qquad i = 1, \ldots, n.$$

Illustrative trials. Coding: 0 = long/lean, 1 = short/rich. Reward: 0 or 1.

- Past experience can affect both **what is chosen** and **how quickly**.

Learning: updating expected reward

For now: we set response times (RT) aside and focus on learning from choices and rewards.

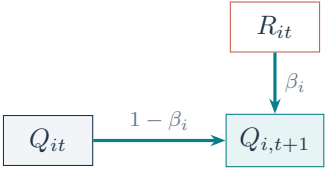
$Q_{it}(a, s)$ is the participant's current value for action a in state s.

$$Q_{i,t+1}(a, s) = Q_{it}(a, s) + \underbrace{\beta_i}_{\text{learning rate}} \underbrace{\{R_{it} - Q_{it}(a, s)\}}_{\text{prediction error}}.$$

Update the visited pair $(a, s) = (A_{it}, S_{it})$. Other values stay unchanged.

$$Q_{i,t+1} = (1 - \beta_i)Q_{it} + \beta_i R_{it}$$

Small β_i: long memory.
Large β_i: more weight on the latest reward.



Prediction-error learning / Rescorla–Wagner update. Huys et al. (2013).

Decision-making: from values to choices

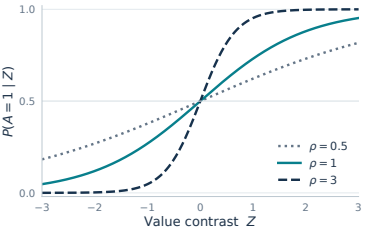
Value contrast

$$Z_{it} = Q_{it}(1, S_{it}) - Q_{it}(0, S_{it}).$$

Standard choice rule

$$\text{logit}\, P(A_{it} = 1 \mid Z_{it}) = \rho_i Z_{it}.$$

Larger $\rho_i > 0$: stronger preference for the action with higher expected reward.

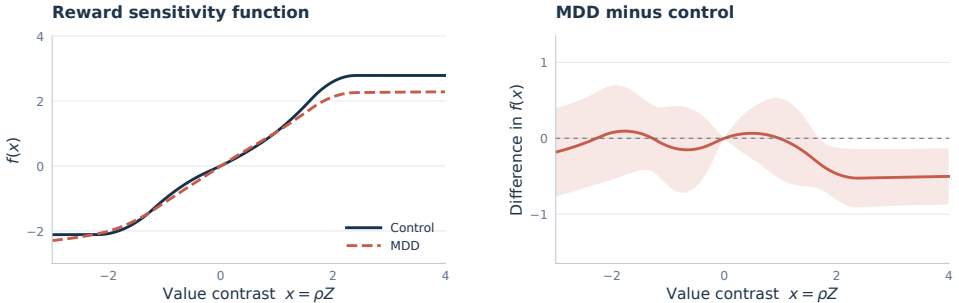


Guo, Zeng, and Wang (2024), JASA

$$\text{logit}\, P(A_{it} = 1 \mid Z_{it}) = f(\rho_i Z_{it}).$$

Learn the response shape from data. f is unknown and nondecreasing, with $f(0) = 0$. Monotone splines allow the response to **flatten or steepen** across value contrasts.

MDD differs at large positive value contrasts



Reward sensitivity: nonlinear, with floor and ceiling effects.

Learning rate: the group difference was not statistically significant.

EMBARC results. Guo, Zeng, and Wang (2024), JASA. Shading shows the uncertainty band for the difference.

Three rules for reward-guided choice

Write $x = \rho_i Z_{it}$.

1. Standard RL

$$p_{\mathrm{RL}}(x) = (1 + e^{-x})^{-1}.$$

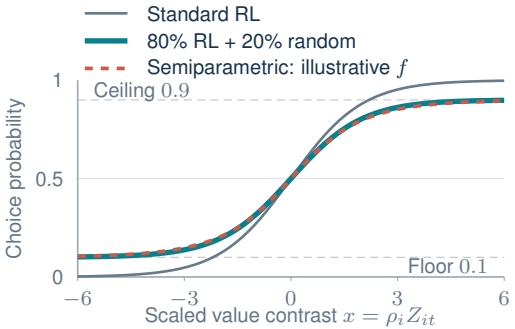
2. RL + random choices

$$p_{\mathrm{mix}}(x) = 0.8\,(1 + e^{-x})^{-1} + 0.2 \times 0.5.$$

3. Semiparametric choice rule

$$p_{\mathrm{SP}}(x) = (1 + e^{-f(x)})^{-1}.$$

f is upper and lower bounded such that $0.1 \leq p_{\mathrm{SP}}(x) \leq 0.9$.

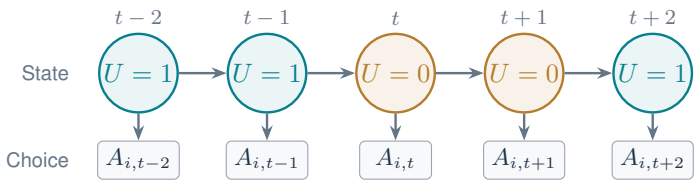


- Semiparametric choice rule with **floor-and-ceiling effect** is consistent with **mixed strategies**.

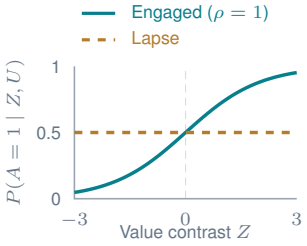
RL-HMM: switching between decision strategies

Guo, Zeng, and Wang (2025), Biostatistics.

A hidden state U_{it} selects the decision rule on each trial.



$$P(U_{it} = k \mid U_{i,t-1} = j) = C_{jk}(t).$$



Engaged $U_{it} = 1$

$$P(A_{it} = 1 \mid U_{it} = 1, Z_{it}) = \frac{1}{1 + e^{-\rho Z_{it}}}.$$

Learned values guide choices.

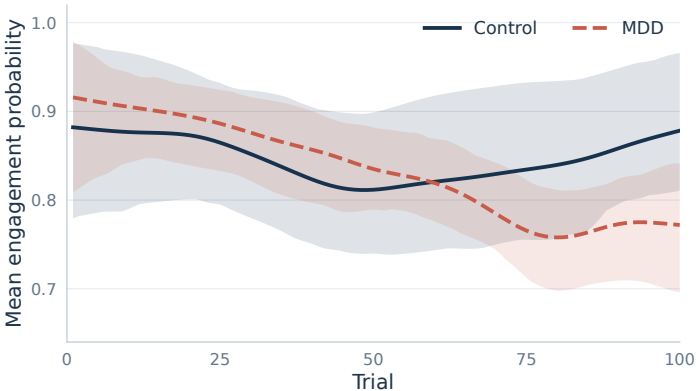
Lapse $U_{it} = 0$

$$P(A_{it} = 1 \mid U_{it} = 0, Z_{it}) = \frac{1}{2}.$$

Random choice, independent of Z_{it}.

Generalization: the framework supports other decision rules and $K > 2$ latent strategies.

Engagement patterns diverge later in the task



Individual engagement

$$H_i(t) = P(U_{it} = 1 \mid \mathcal{D}_i),$$

This is the estimated probability that subject i is engaged at trial t.

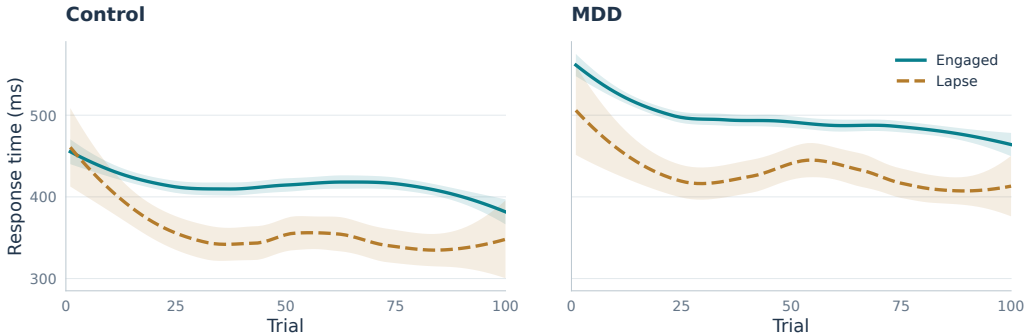
Group engagement

$$\bar{H}(t) = \frac{1}{n} \sum_{i=1}^{n} H_i(t).$$

This estimates the proportion of subjects engaged at trial t.

The fitted MDD trajectory is lower in later trials, suggesting a difference in sustained engagement.

Response times separate the inferred strategies

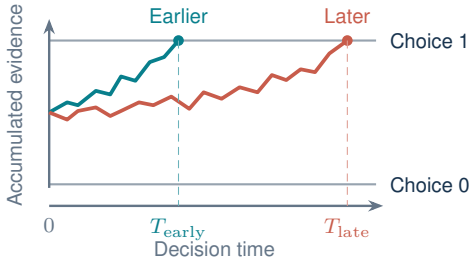


Trials classified as **engaged** generally take longer than **lapse** trials.

Within each strategy, the control group responds faster than the MDD group in these fitted curves.

Why model response time jointly?

Same choice, different decision times



RT also includes encoding and motor time.

A fast response can reflect

Stronger evidence
One option is clearly favored.

Less caution
Less evidence is required before responding.

A different strategy
For example, a rapid guess.

Joint model $p\big(A_{it}, \mathrm{RT}_{it} \mid S_{it}, \mathcal{H}_{it}\big).$

Choice probabilities and RT distributions **jointly constrain the decision process**.

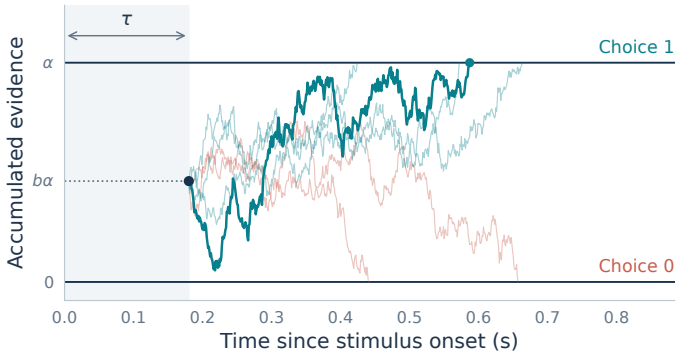
Schematic trajectories, not fitted data. Ratcliff (1978); Pedersen, Frank, and Biele (2017); Bian, Guo, and Wang (2025).

Drift diffusion as evidence accumulation

$$\mathrm{d}X(u) = v\,\mathrm{d}u + \sigma\,\mathrm{d}W(u),$$
$$X(0) = b\alpha.$$

- v drift rate
- α boundary separation
- b initial bias
- τ non-decision time

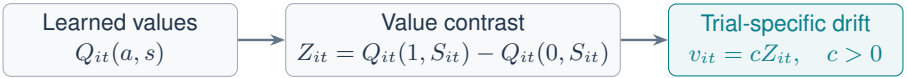
$$\mathrm{RT} = \tau + T_{\mathrm{hit}}.$$



A boundary hit determines the choice.

Simulated illustration. Ratcliff (1978). Here u is time within a trial; t indexes trials. Fix $\sigma = 1$ in the fitted model.

RL-DDM links learned values to the drift rate



$$(\mathrm{RT}_{it}, A_{it}) \mid S_{it}, \mathcal{H}_{it} \sim \mathrm{WFPT}(\alpha, b, v_{it}, \tau).$$

The Wiener first-passage time (WFPT) distribution joint density

Let $d = r - \tau > 0$ and write $v = v_{it}$. **Choice** 1 hits the upper boundary α.

$$p(r, 1 \mid \alpha, b, v, \tau) = \frac{\pi}{\alpha^2} e^{\,v\alpha(1-b) - v^2 d/2} \sum_{k=1}^{\infty} k \sin\big(k\pi(1-b)\big) e^{-k^2\pi^2 d/(2\alpha^2)}.$$

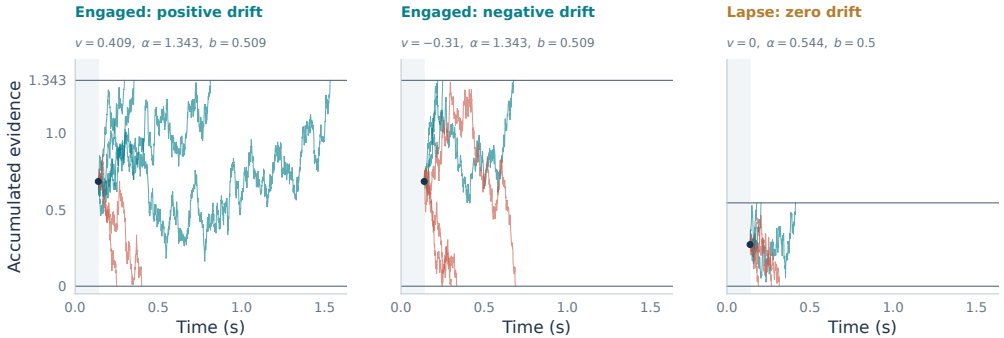
Choice 0: reflect the drift and starting point.

$$p(r, 0 \mid \alpha, b, v, \tau) = p(r, 1 \mid \alpha, 1-b, -v, \tau).$$

Zero for $r \le \tau$; **integrating over RT gives the probability of each choice**.

WFPT: Wiener first-passage time distribution. Navarro and Fuss (2009); Pedersen, Frank, and Biele (2017).

Three decision regimes



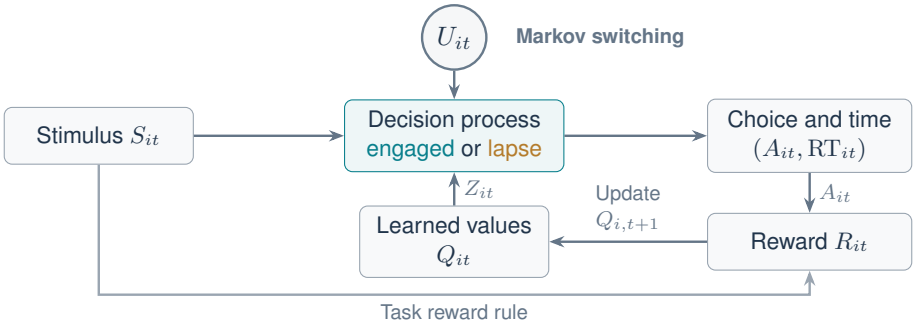
The lapse model uses zero drift, no initial bias, and a separate boundary.

Its smaller fitted boundary permits faster decisions.

Simulated paths with fitted parameters from Bian et al. (2025), $\tau = 0.139$ s, and $\sigma = 1$. Teal paths hit choice 1; coral paths hit choice 0.

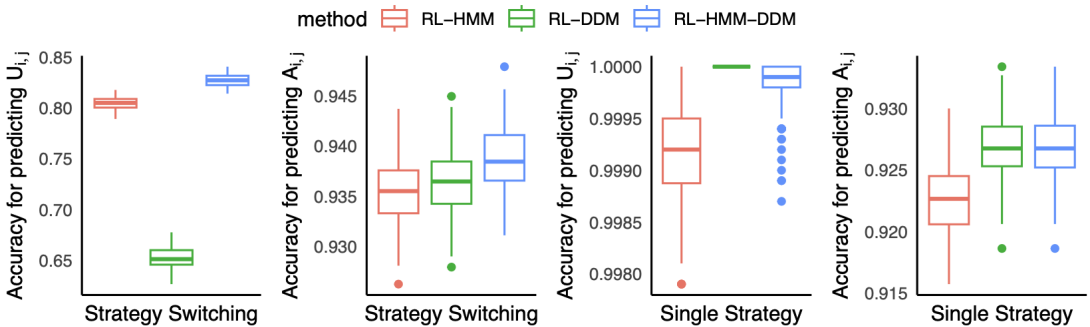
RL-HMM-DDM joins switching, choice, and time

Bian, Guo, and Wang (2025), AOAS.



$$(\mathrm{RT}_{it}, A_{it}) \mid U_{it}, S_{it}, \mathcal{H}_{it} \sim \begin{cases} \mathrm{WFPT}(\alpha_1, b, cZ_{it}, \tau), & U_{it} = 1: \text{ engaged}, \\ \mathrm{WFPT}(\alpha_0, \frac{1}{2}, 0, \tau), & U_{it} = 0: \text{ lapse}. \end{cases}$$

Simulation: recovering strategies and choices



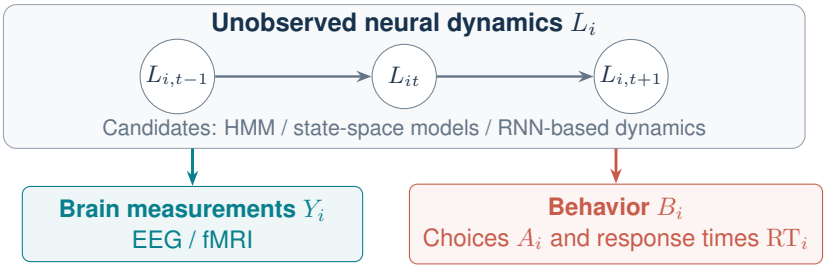
With switching: RL-HMM-DDM improves strategy recovery and choice prediction.

Without switching: its choice accuracy is comparable to RL-DDM.

Bian, Guo, and Wang (2025). Boxplots show accuracy for latent strategy U and choice A across simulation replicates.

Ongoing work: linking brain and behavior

Goal: infer shared neural dynamics from brain measurements and behavior.



General generative model Given task inputs S_i

$$p_\theta(L_i, Y_i, B_i \mid S_i) = \underbrace{p_\theta(L_i \mid S_i)}_{\text{latent dynamics}} \times \underbrace{p_\theta(Y_i \mid L_i, S_i)}_{\text{brain model}} \times \underbrace{p_\theta(B_i \mid L_i, S_i)}_{\text{behavior model}}.$$

Subject-level trajectories; shared-state assumption: $Y_i \perp B_i \mid L_i, S_i$.

Learning, strategy, and timing tell different parts of the story.

Flexible response functions reveal where groups differ in reward sensitivity.

Latent-state models capture shifts in decision strategies across trials.

Joint models integrate choices, response times, and brain signals.

Thank you

Xingche Guo xingche.guo@uconn.edu

Joint work with Yuan Bian, Yuanjia Wang, and, Donglin Zeng.

References: the modeling sequence

Guo, X., Zeng, D., and Wang, Y. (2024). A Semiparametric Inverse Reinforcement Learning Approach to Characterize Decision Making for Mental Disorders. *JASA*, 119(545), 27–38.

doi:10.1080/01621459.2023.2261184

Guo, X., Zeng, D., and Wang, Y. (2025). HMM for discovering decision-making dynamics using reinforcement learning experiments. *Biostatistics*, 26(1), kxae033.

doi:10.1093/biostatistics/kxae033

Bian, Y., Guo, X., and Wang, Y. (2025). Joint modeling for learning decision-making dynamics in behavioral experiments. *Annals of Applied Statistics*, 19(4), 3372–3393.

doi:10.1214/25-AOAS2112

References: tasks and foundational models

Pizzagalli, D. A., Jahn, A. L., and O'Shea, J. P. (2005). Toward an objective characterization of an anhedonic phenotype: a signal-detection approach. *Biological Psychiatry*, 57, 319–327.

Trivedi, M. H., et al. (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): rationale and design. *Journal of Psychiatric Research*, 78, 11–23.

Huys, Q. J. M., Pizzagalli, D. A., Bogdan, R., and Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biology of Mood & Anxiety Disorders*, 3, 12.

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

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

Ashwood, Z. C., et al. (2022). Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25, 201–212.