#DRA #human-memory #slot-machines

Slot machines pilot v3.1

Outline

- Evaluate performance of all ($N = 61$) subjects
  - Split by chance-level performance
    - see Methods > Determining chance-level performance
  - Fit hierarchical models to each populations' choice data
    - see Methods > Hierarchical Bayesian logistic regression
- Chance-level subjects ($N = 29$):
  - Verify that the posterior over psychometric slopes are indeed centered roughly around zero at group level & individual
    - see Appendix > Bad subjects
- Above-chance subjects ($N = 32$):
  - Split by model-type by testing posterior over psychometric slopes (see Methods > Testing model signatures)
    - Classify subjects into the following model types:
      - DRA ($N = 15$)
      - Freq ($N = 9$)
      - Stakes ($N = 2$)
      - Equal Precision ($N = 6$)
    - Fit hierarchical models to each of these sub-populations
    - See Results > Classifying subjects
  - Use psytrack (Roy, Pillow, et al.) to see trends through learning

Model-fitting & Bayesian model comparison

For the immediate future, the plan is to fit the data to each of the models, i.e. compute the log-likelihood of the data for each participant.

- The plan is to use Luigi's Variational Bayes Monte Carlo package for posterior and model inference.
- In addition to the four models we have, we will include the standard RL model as well as a Bayesian ideal observer model.

Methods

Determining chance-level performance

First, we simulate a random agent for the task experienced by the humans (with the same number of trials). We repeat this 100,000 times to get a null distribution for performance (for definition, see box below). We can then filter the subjects into two groups by comparing their performance against the 95th quantile of the null distribution (one-tailed test).

Definition of performance

Performance is a (normalized) scale from -100 to 100 representing the mean expected reward over all trials in the test blocks, where expected reward for each trial is what a subject might have received if the slot machines were not noisy.

- A score of 100 on this scale would represent the maximum expected reward.
- A score of 0 would represent chance-level expected reward for our task (or 50% choice accuracy).
- A score of -100 would be highly unexpected but possible if an agent chose the wrong option for every single trial.

Alternative null distributions

We also computed a null distribution based on choice accuracy instead of performance, and ones by shuffling the responses of the participants instead of simulating a random agent to account for biases. Neither of these changed the filtering results by more than $N = 1$ subject from $N = 61$ total subjects.

Hierarchical Bayesian logistic regression

We construct a Bayesian logistic regression model that jointly models the choice probability for all of the four slot machines, each one with its own slope and intercept parameter. Our model can be written out as:

$$p\left(y_t | g\left(\alpha_0 + \sum_s \mathbb{I}_{SM_t=s}\left(\alpha_s + x_{t,s}\beta_s\right)\right), \theta\right)$$

where:

- $p(y_t|g(\cdot),\theta) : bernoulli$
- $g : logit^{-1}$
- $s \in \{1, 2, 3, 4\}$ indexes the slot machine

- t indexes the trial
- $\mathbb{I}_{SM=s}(\cdot)$ is an indicator function which acts as the identity function when the slot machine on that trials is s
- $y_t \in \{0, 1\}$ is the boolean response
- $x_{t,s} \in \mathbb{R}$ is the expected reward slot machines s
  - This is what we used to call "difficulty", but "expected reward" makes more sense
  - In our current task, the values only take one of $\{-2\delta, -\delta, \delta, 2\delta\}$, but in theory, they could be anything in $\mathbb{R}$.

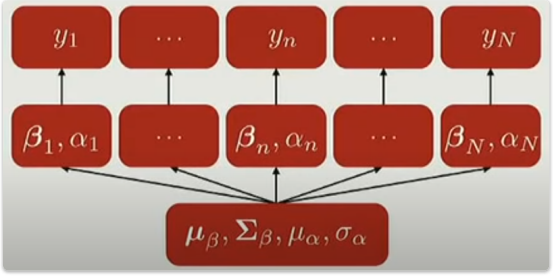
Individual subjects' parameters:

- are drawn from the group-level priors
- $\beta_n \sim \mathcal{N}(\mu_\beta, \Sigma_\beta) \in \mathbb{R}^4$
  - slope of the psychometric curve for participant n
  - 4D vector, one dim for each slot machine s
- $\alpha_n \sim \mathcal{N}(\mu_\alpha, \Sigma_\alpha) \in \mathbb{R}^4$
  - bias for participant n
  - 4D vector, one dim for each slot machine s
- $\alpha_{0_n} \sim \mathcal{N}(\mu_{\alpha_0}, \sigma_{\alpha_0}) \in \mathbb{R}$
  - baseline bias for participant n
  - this parameter is redundant and unnecessary

So far, we haven't talked at all about the hierarchy. However, the model is hierarchical, meaning that the subject-level parameters are drawn from group-level priors, which are themselves drawn from reasonable hyper-priors. The group-level priors are:

- $\theta_{\text{group}} = \{\mu_{\alpha_0}, \sigma_{\alpha_0}, \mu_\beta, \Sigma_\beta, \mu_\alpha, \Sigma_\alpha\}$

By doing this, individual subjects' parameters inform the group-level parameters and vice versa. The hierarchical model can be seen with the following schematic:



(we have ignored the α_0 parameter here, and the $\alpha_i \ \ \forall i \in \{1, \ldots, N\}$ as well as μ_α should be **bold** since they are 4D vectors; will update soon)

Testing model signatures

The hierarchical model fits give us the posterior over psychometric slopes, $\beta \in \mathbb{R}^4$. We take 100,000 samples from these 4D posteriors to test for the model signatures as per Antonio's original proposal of the basic eye test. We pose multiple Boolean questions to each sample, one set for each for the four models we test - DRA, Frequency, Stakes, and Equal Precision - and see how many of the 100,000 samples pass the whole set of questions combined with the AND operator, i.e. $q_1 \ \& \ q_2 \ \& \ \dots \ q_n$, where q_i is the i^{th} question in the set. This gives us a crude approximation of what we want, but at the moment, this is the best we can get.

Before specifying the questions in each set, here's a refresher on how the four slot machines are factorized in the task:

	High stakes	Low stakes
High freq	1	2
Low freq	3	4

The sets of Boolean questions that we pose for each of the models is as follows:

DRAx	DRA+	Freq	Stakes	EP
$\beta_1 > \beta_2$	$\beta_1 \geq \beta_2$	$\beta_1 = \beta_2$	$\beta_1 = \beta_3$	$\beta_1 = \beta_2$
$\beta_1 > \beta_3$	$\beta_1 \geq \beta_3$	$\beta_3 = \beta_4$	$\beta_2 = \beta_4$	$\beta_1 = \beta_3$
$\beta_1 > \beta_4$	$\beta_1 \geq \beta_4$	$\beta_1 > \beta_3$	$\beta_1 > \beta_2$	$\beta_1 = \beta_4$
$\beta_2 > \beta_4$	$\beta_2 \geq \beta_4$	$\beta_1 > \beta_4$	$\beta_1 > \beta_4$	$\beta_2 = \beta_4$
$\beta_3 > \beta_4$	$\beta_3 \geq \beta_4$	$\beta_2 > \beta_3$	$\beta_3 > \beta_2$	$\beta_2 = \beta_4$
	not Freq	$\beta_2 > \beta_4$	$\beta_3 > \beta_4$	$\beta_3 = \beta_4$
	not Stakes			
	not EP			

For DRA, we have two different tests. One of them (DRAx) is based on strict ordering of the slopes, which is know from simulations that the model doesn't always respect. The other (DRA+) is based on what we know from simulations with the difference being that the slopes can be greater than or equal to the next one in the order. Since this would be a superset which includes all the other three models (Freq, Stakes, EP), we explicitly add three more Boolean question to see that it is *not* one of the other three models.

Test for equality

We test $\beta_i = \beta_j$ with a threshold ϵ as $|\beta_i - \beta_j| < \epsilon$.

The problem is that the threshold, ϵ, is arbitrary. In practice, we determine it by first computing the median standard deviation, σ_μ, of the marginal posteriors $\beta_1, \ldots, \beta_4$, then setting $\epsilon = 2\sigma_\mu$.

⚠ Caveat

Testing for equality selects a part of 4D space as a hypercube. On the other hand, testing for inequality splits the the space in half (per dimension). This means that there is *a lot more* space where points could lie for inequality as opposed to the equality test, which may be problematic.

Fix

In practice, this only really affects the tests at low thresholds ϵ when testing for equality. We were previously using $\epsilon = \sigma_\mu$ and have since switched to $\epsilon = 2\sigma_\mu$ which seems to work much better. We validate our test results later with sanity checks that involve building separate hierarchical models for each group of subjects classified with this test.

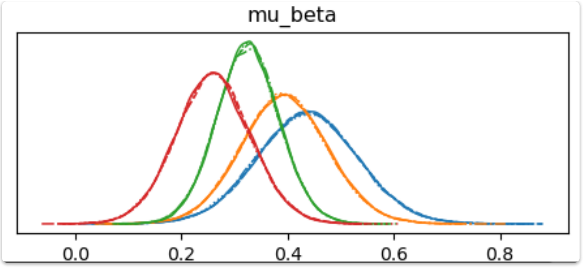
Psytrack

Psytrack is a method to fit dynamical psychophysical models to behavioral data developed by Nick Roy from Jonathan Pillow's lab in Princeton. Instead of assuming the same strategy throughout learning, psytrack fits a posterior over the slope of the psychometric curve dynamically through time, allowing us to visualize trends over the course of learning. For more details, refer to their Neuron paper and their Github repository.

The one caveat is that this method only works for individual subjects unlike our hierarchical model.

Results

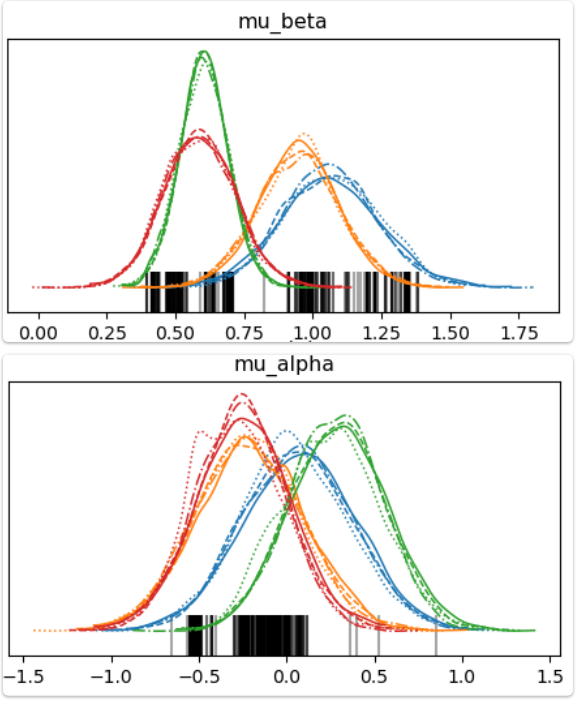
First, we fit the hierarchical model for all $N = 61$ subjects. Here's the group-level posterior over β, the slope of psychometric curves.



Sorry for the missing legends, but the colors represent the four slot machines: 1 = blue, 2 = yellow, 3 = green, 4 = red.

After evaluating their performance, we find that $N = 29$ perform at or below chance and $N = 32$ are above chance. See Appendix > Performance scores for the full table of accuracy and performance scores for all participants.

We fit hierarchical models to each group, and verify that the posterior over β is indeed centered around zero for the subjects performing at or below chance (see Appendix > Bad subjects). The good subjects' group-level posteriors over β (slope) and α (bias) suggest that subjects may be following DRA or Frequency-based strategies:



To see the posterior over β for the individual subjects ordered by performance and also how β correlates with performance, see Appendix > Good subjects.

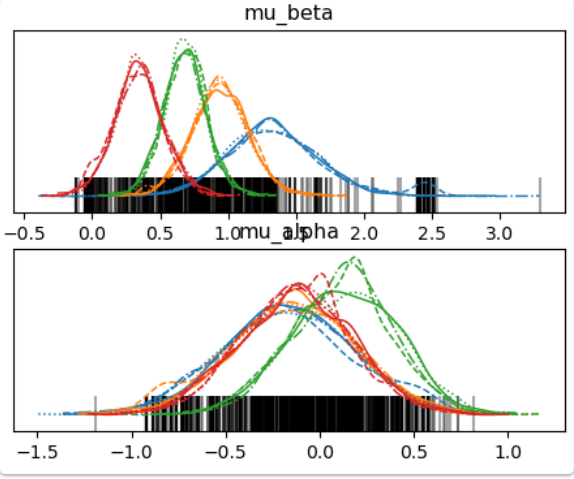
Classifying subjects

Next, we classify subjects into four groups. We do this by running the tests described in Methods > Testing model signatures. This gives us the following four groups. The results here should come as no surprise since it is by design that we

see these trends. If anything, this only serves as a sanity check to validate our testing procedure.

DRA $(N = 15)$

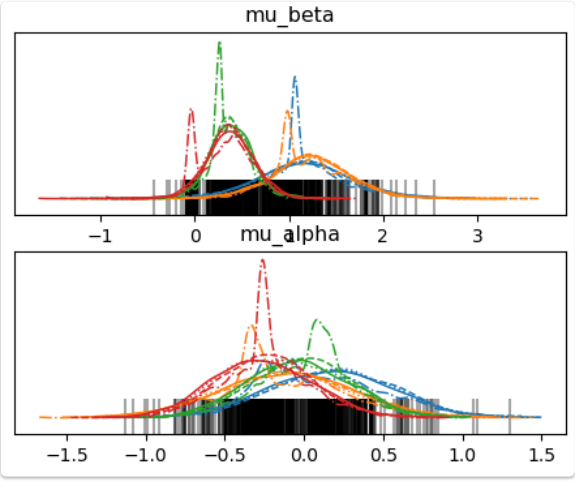
Here are the group-level posteriors over β (top, slope) and α (bottom, bias):



(β_1 blue) > (β_2 yellow) > (β_3 green) > (β_4 red)

Freq $(N = 9)$

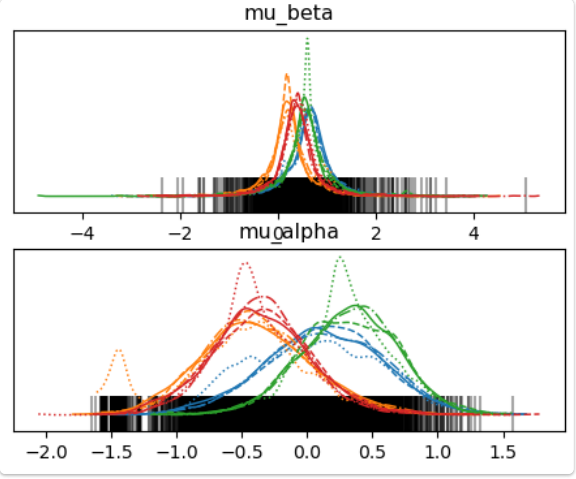
Here are the group-level posteriors over β (top, slope) and α (bottom, bias):



(β_1 blue) $\approx$ (β_2 yellow) > (β_3 green) $\approx$ (β_4 red)

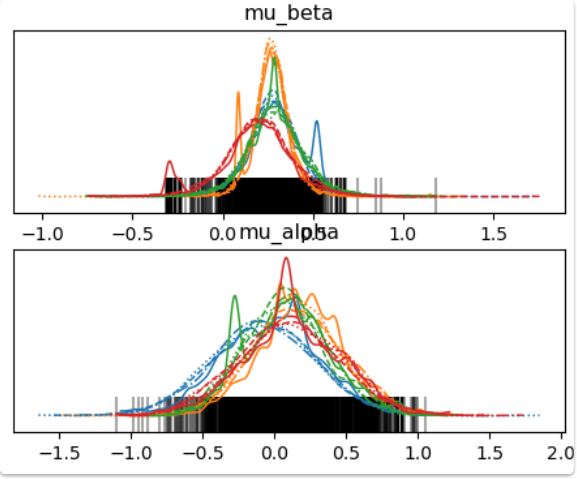
Stakes $(N = 2)$

Here, it's really unclear and seems like the biases may be driving it more than the slopes. But in any case, it's just 2 participants. Who knows what's happening.



Equal Precision $(N = 6)$

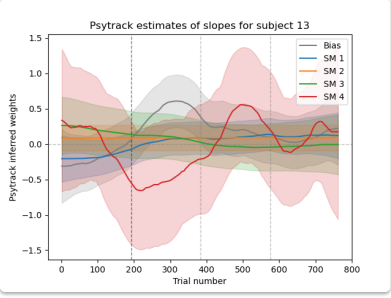
These subjects also seem to be correctly classified.

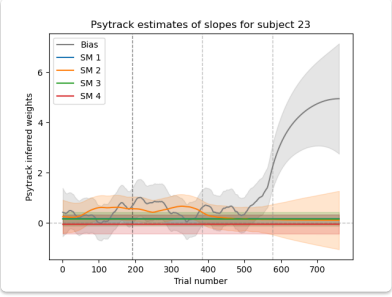


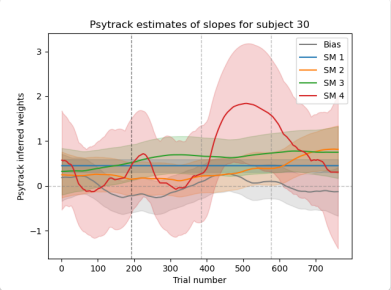
Psytrack trends in learning

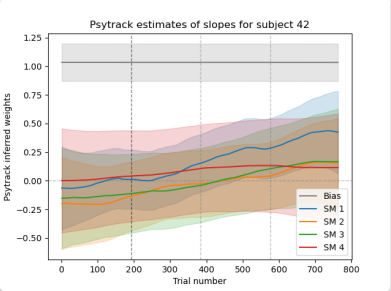
We didn't notice something striking with the learning dynamics, but here are the results for completeness. The one caveat with the current analyses is that psytrack seems to have a failure mode which sometimes leads to a flat line throughout learning. It's likely that this is because of the very limited number of trials we have for the infrequent slot machines (96 per subject).

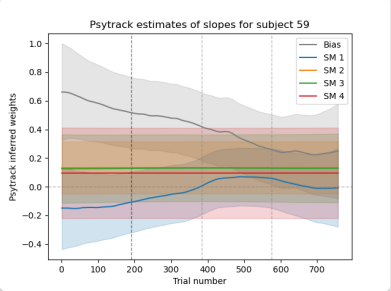
Equal Precision examples



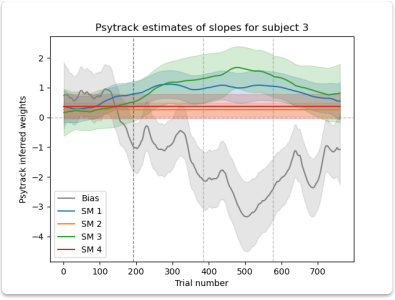


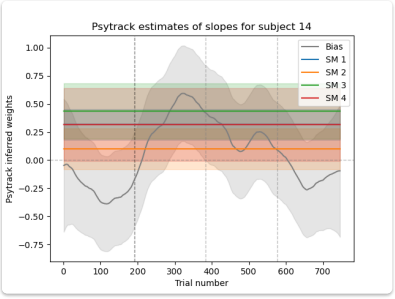




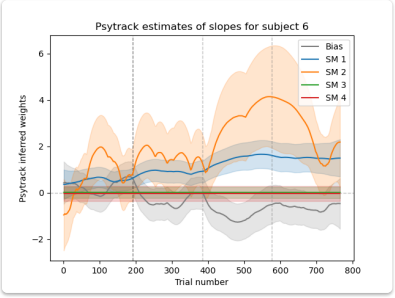


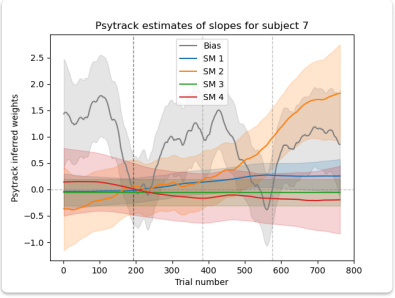
Stakes examples

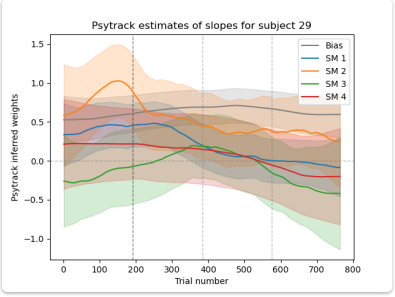


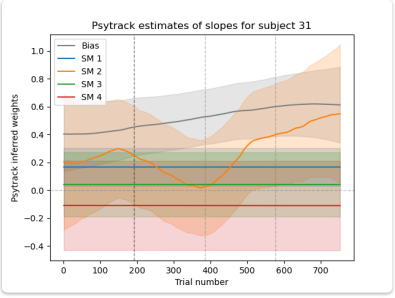


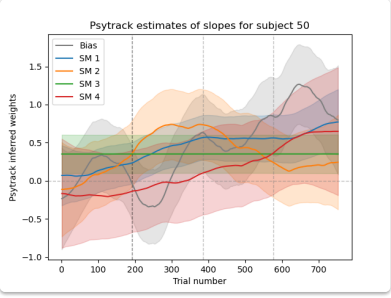
Freq examples

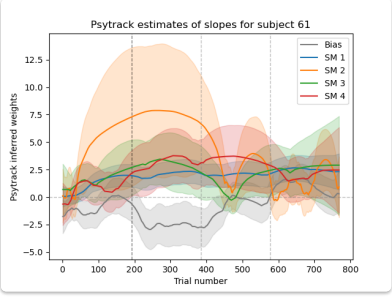




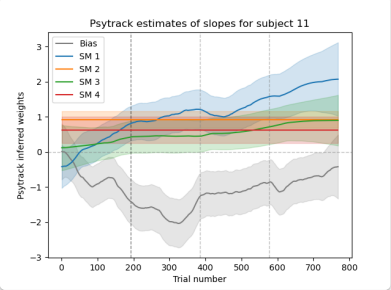


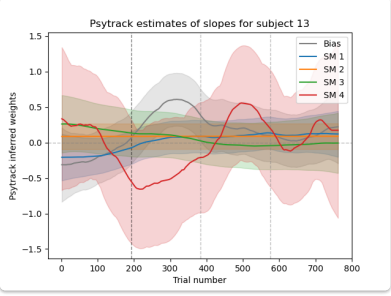


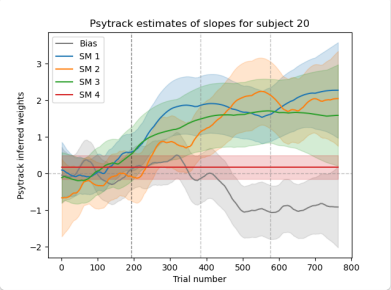

Psytrack estimates of slopes for subject 50
Bias
SM 1
SM 2
SM 3
SM 4
Psytrack inferred weights
Trial number

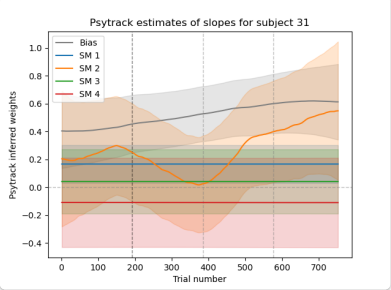

Psytrack estimates of slopes for subject 61
Bias
SM 1
SM 2
SM 3
SM 4
Psytrack inferred weights
Trial number

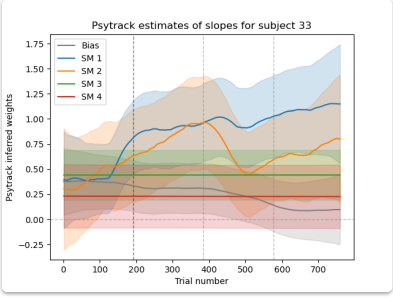
DRA

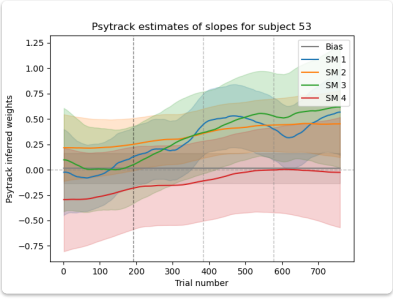

Psytrack estimates of slopes for subject 11
Bias
SM 1
SM 2
SM 3
SM 4
Psytrack inferred weights
Trial number

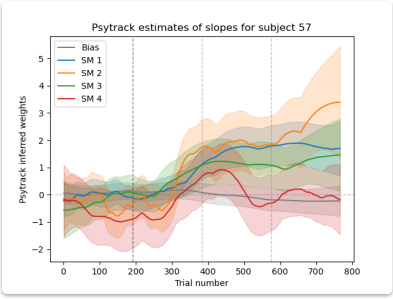

Psytrack estimates of slopes for subject 13
Bias
SM 1
SM 2
SM 3
SM 4
Psytrack inferred weights
Trial number


Psytrack estimates of slopes for subject 20
Bias
SM 1
SM 2
SM 3
SM 4
Psytrack inferred weights
Trial number


Psytrack estimates of slopes for subject 31
Bias
SM 1
SM 2
SM 3
SM 4
Psytrack inferred weights
Trial number







Appendix

Performance scores

	id	performance	accuracy	above_chance
0	21	90.74	94.62	True
1	61	84.03	89.76	True
2	48	73.30	80.73	True
3	20	71.67	81.60	True
4	36	65.74	81.25	True
5	46	65.27	77.60	True
6	56	63.77	79.51	True
7	37	63.01	78.47	True
8	6	60.50	78.99	True
9	19	59.91	75.00	True
10	57	57.74	77.08	True
11	8	56.64	73.96	True
12	33	54.46	72.74	True
13	11	53.94	76.91	True

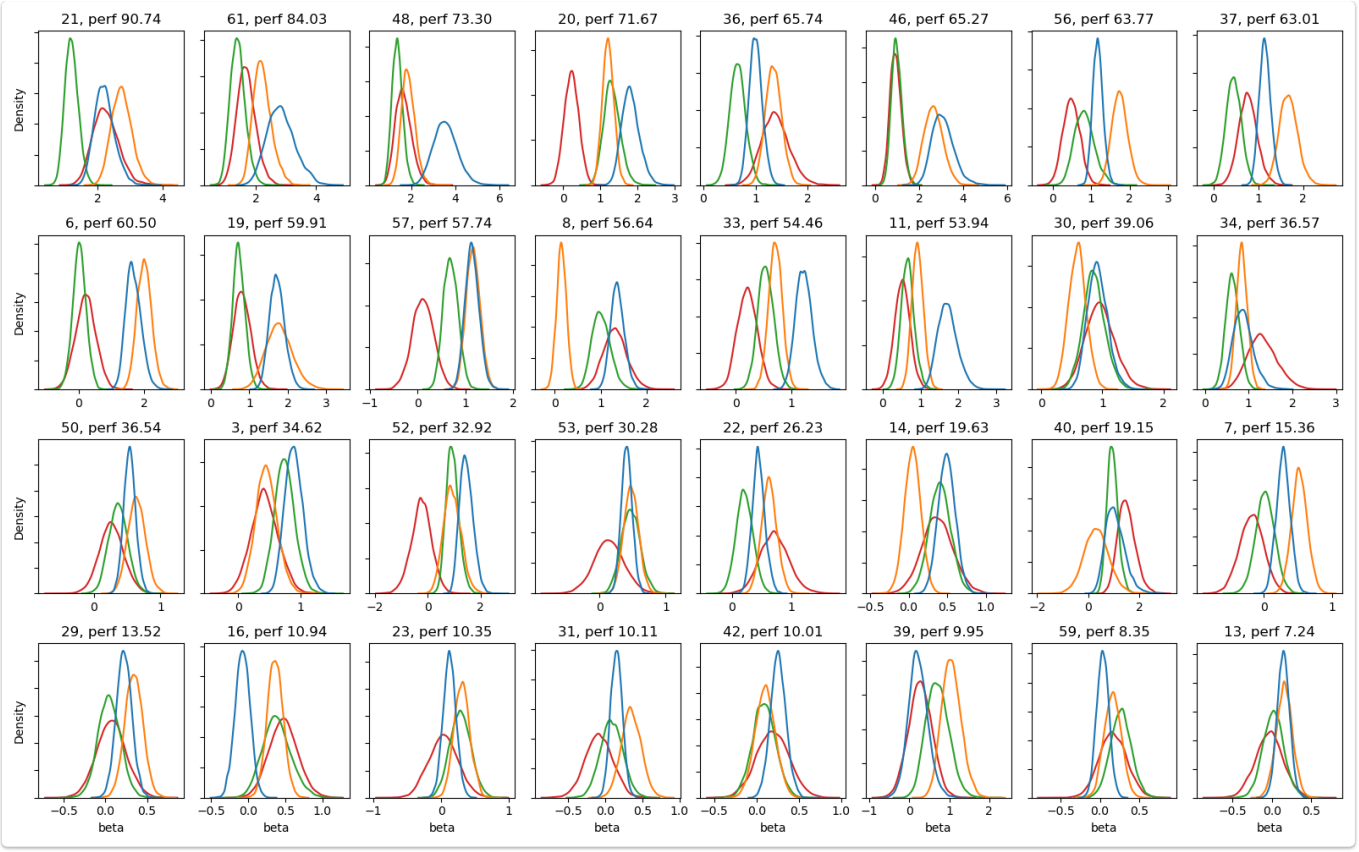
	id	performance	accuracy	above_chance
14	30	39.06	67.71	True
15	34	36.57	68.06	True
16	50	36.54	64.41	True
17	3	34.62	63.19	True
18	52	32.92	62.33	True
19	53	30.28	61.98	True
20	22	26.23	60.76	True
21	14	19.63	56.60	True
22	40	19.15	57.99	True
23	7	15.36	58.16	True
24	29	13.52	56.42	True
25	16	10.94	55.03	True
26	23	10.35	53.47	True
27	31	10.11	54.34	True
28	42	10.01	53.99	True
29	39	9.95	54.34	True
30	59	8.35	54.17	True
31	13	7.24	52.95	True
32	32	6.96	52.60	False
33	44	6.48	53.99	False
34	4	6.15	50.35	False
35	1	5.10	52.43	False
36	26	4.93	51.91	False
37	35	4.33	49.65	False
38	54	3.29	49.83	False
39	49	2.64	48.61	False
40	5	2.09	50.35	False
41	15	1.89	49.31	False
42	45	1.86	50.87	False
43	28	1.86	52.26	False
44	18	1.65	49.65	False
45	47	0.47	48.26	False

	id	performance	accuracy	above_chance
46	41	0.47	50.00	False
47	24	0.43	48.96	False
48	55	-0.93	49.31	False
49	25	-0.93	48.44	False
50	58	-2.31	49.31	False
51	51	-2.55	48.44	False
52	27	-3.06	48.78	False
53	10	-3.28	47.40	False
54	9	-3.60	45.31	False
55	12	-4.22	47.92	False
56	43	-7.43	45.49	False
57	17	-8.59	45.49	False
58	60	-8.62	45.83	False
59	2	-12.04	41.84	False
60	38	-22.96	39.58	False

Good subjects

β ordered by performance

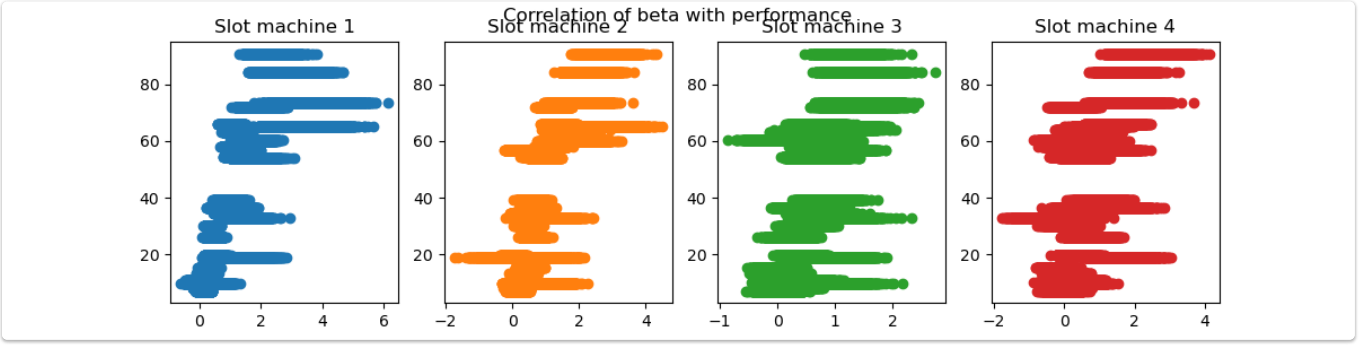
Here's the plot for the posterior over psychometric slopes for the subjects who perform above chance. The plots are ordered by performance with the highest performing subjects in the top row. The key thing to notice is that the mode of the posterior over β decreases as their performance drops. For the last few, who are barely above chance, the posterior over β are centered roughly around 0.



The subjects' ids and performance are listed on top of each panel. The colors represent the four slot machines, 1 = blue, 2 = yellow, 3 = green, 4 = red. Sorry for the missing legend.

Correlation of β with performance

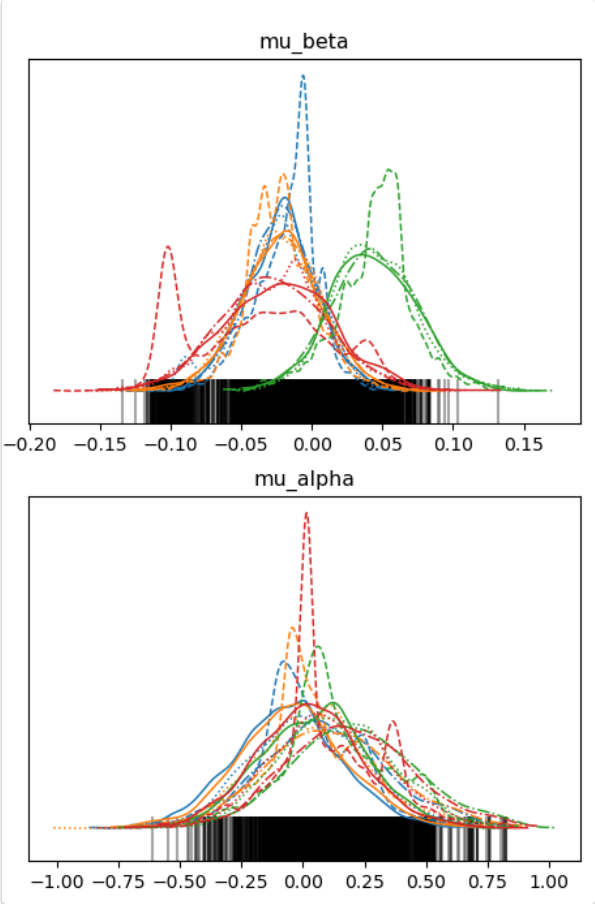
We further demonstrate the point by showing that the performance is highly correlated with the β for each slot machine, which validates our hierarchical model.



Bad subjects

The posterior over β for the subjects performing at or below chance is centered around zero.

Group-level



Individual, ordered by performance