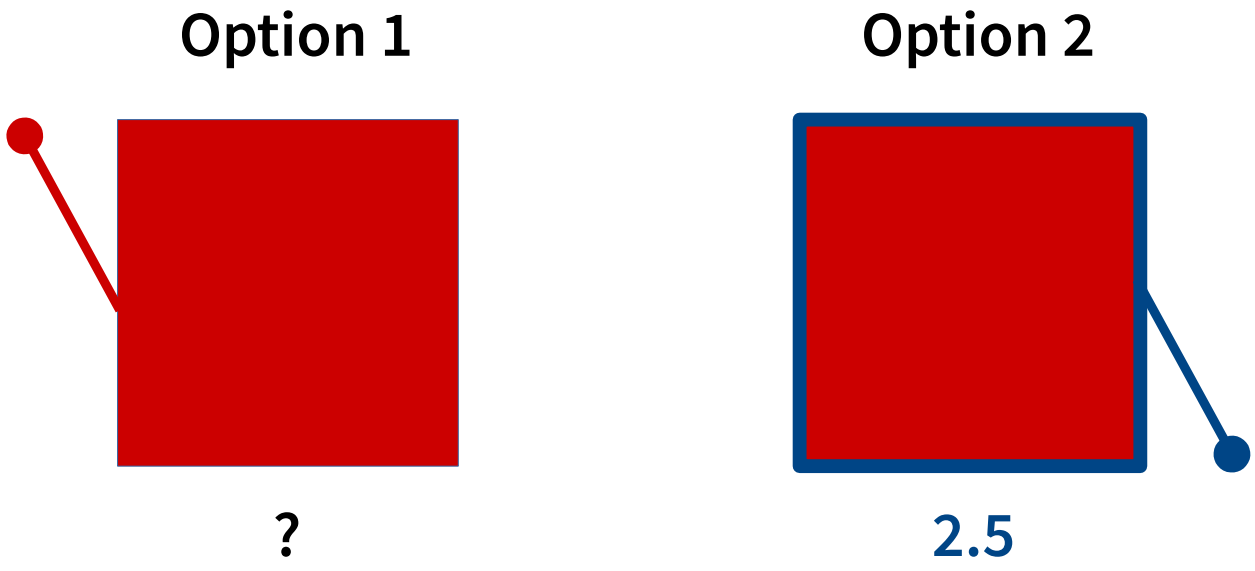
Human behavior in contextual multi-armed bandits

Hrvoje Stojic,
Pantelis Pipergias Analytis and Maarten Speekenbrink

Pompeu Fabra University,
Max Planck Institute for Human Development, University College London

August 20, 2015

Multi-armed bandit (MAB) problem



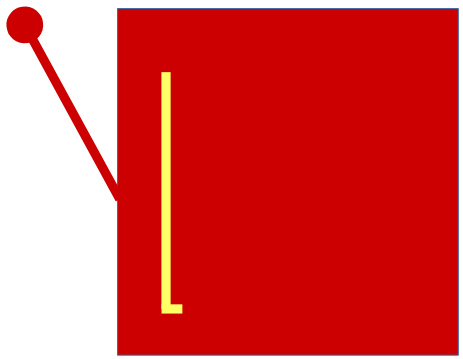
$$N(\mu_1, \sigma_1) \qquad N(\mu_2, \sigma_2)$$

Blind product testing

Realistic decision problem...

Contextual multi-armed bandit (CMAB) problem

Option 1

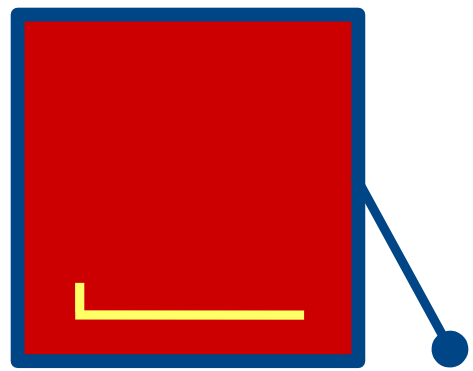


?

$$N(f(\cdot), \sigma_1)$$

$$N(\omega_1 x_1 + \omega_2 x_2, \sigma)$$

Option 2



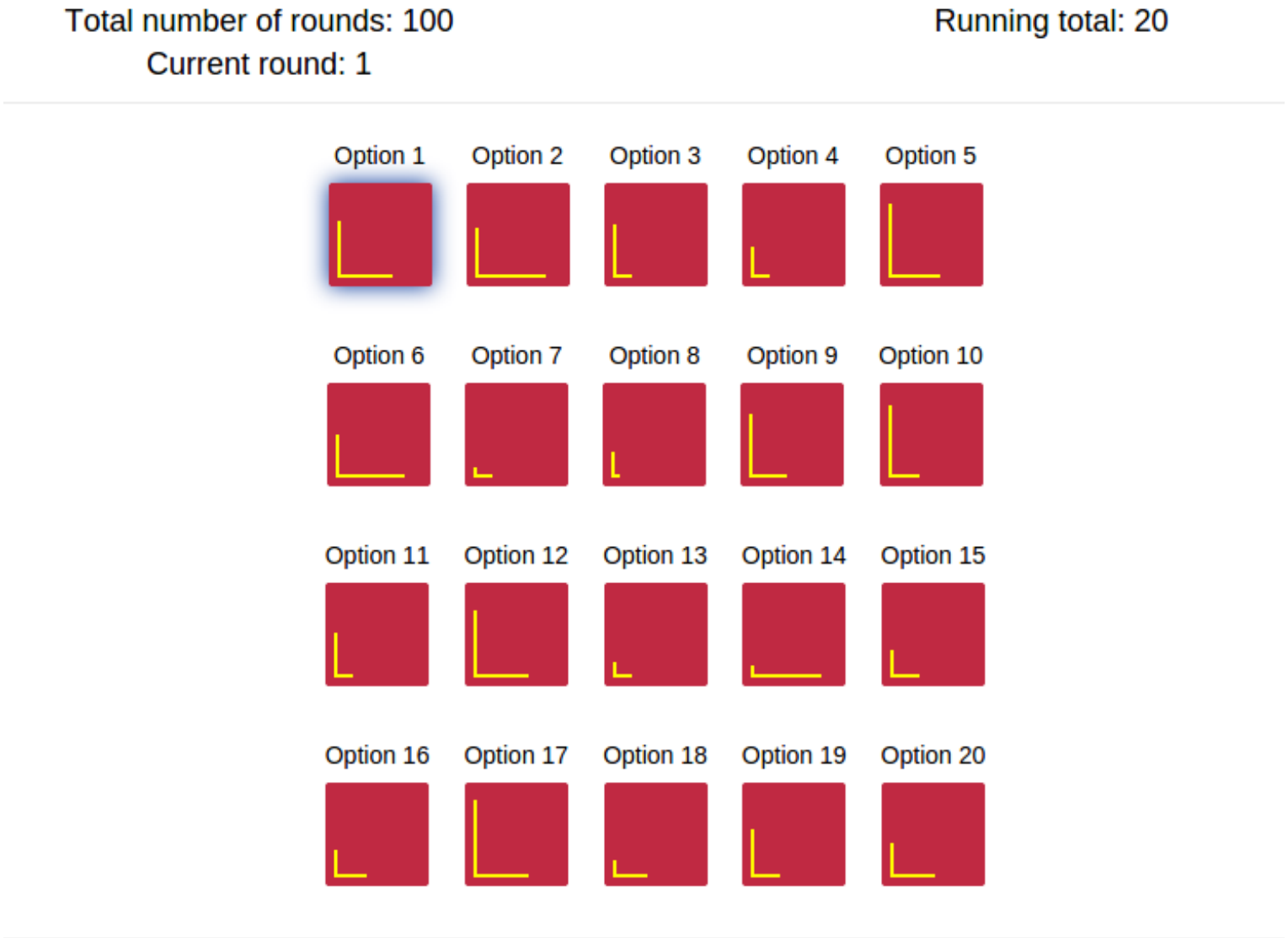
2.5

$$N(f(\cdot), \sigma_2)$$

$$N(\omega_1 x_1 + \omega_2 x_2, \sigma)$$

Really realistic decision problem...

Our CMAB task



Why is CMAB problem interesting?

Highlights:

1. Closer to realistic situations rich in features.
2. $TD(\lambda)$ heavily affected by the curse of dimensionality - structure learning might be a solution.
3. Function learning changes the exploration-exploitation problem!
4. We can study generalization, transfer of learning, novelty...

Goal of this study:

1. Provide characterization of human behavior in CMAB problems.
2. Develop function learning based reinforcement learning models to explain the behavior.

Experimental Design

Experiment 1 - High noise

- **Training phase** – between subject design:
  - Contextual multi-armed bandit (CMAB) task – two informative features are visually displayed
  - Classic multi-armed bandit (MAB) task – control group, features are not visible
  - 20 alternatives, 100 trials
- **Test phase** – one shot choices between 3 arms in 70 trials, without outcome feedback
- 145 participants – Amazon Turk – monetary payoffs

Experiment 2 - Low noise

- **Reduced** standard deviation of the error term
- 143 participants – Amazon Turk – monetary payoffs

Experimental Task

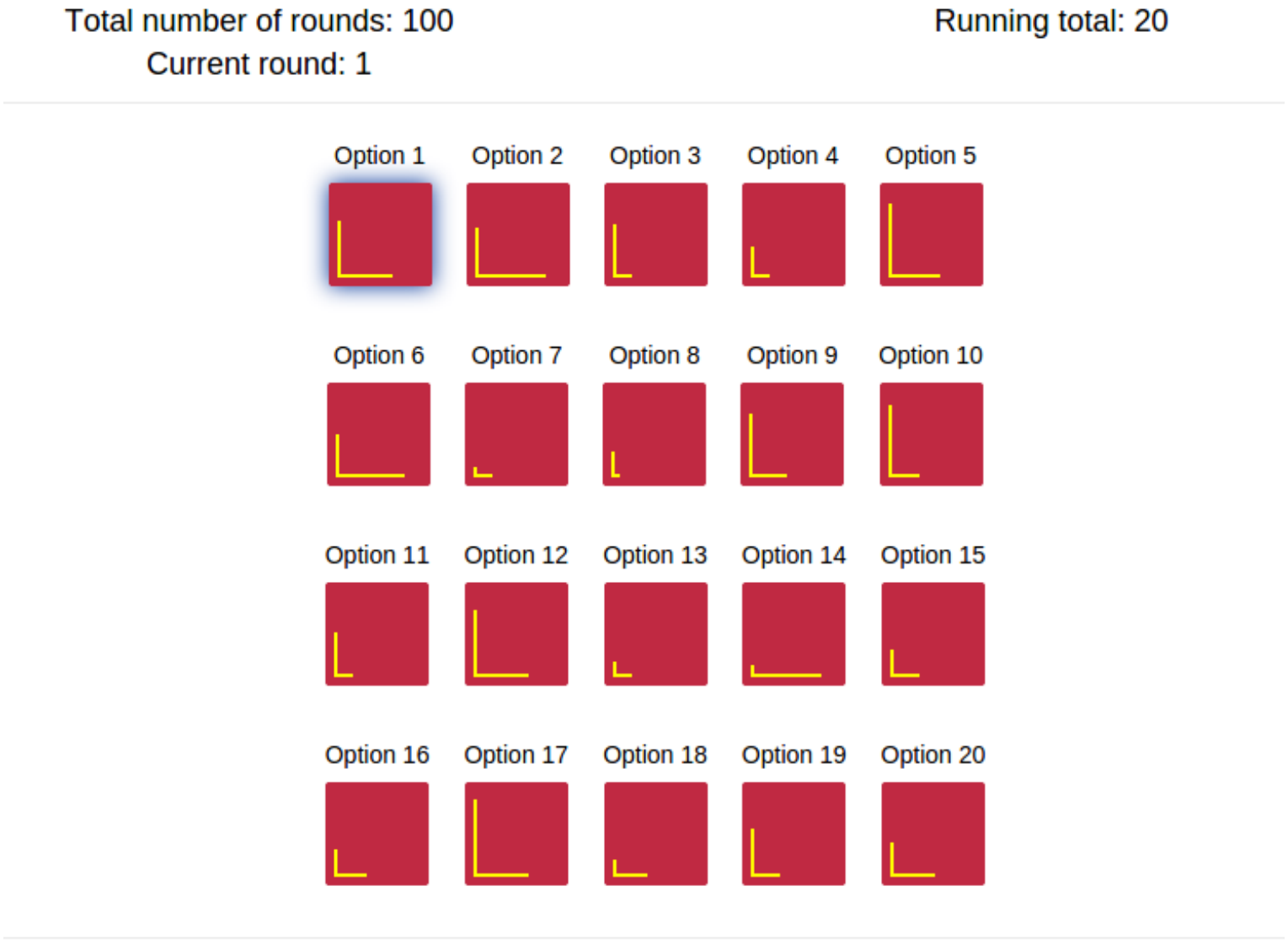
The task in all conditions

- For each arm j in trial t, the payoffs $R_j(t)$ were computed as:

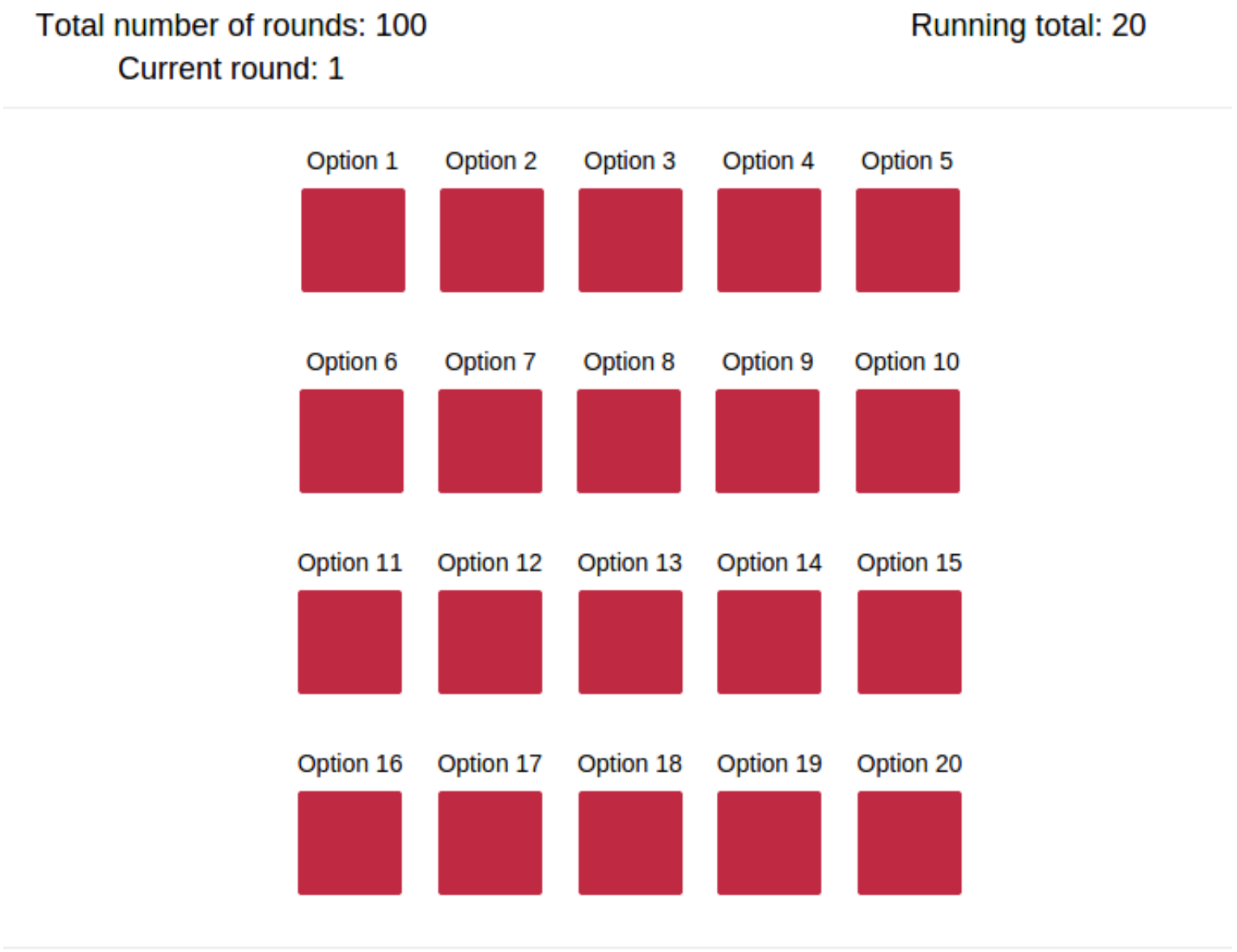
$$R_j(t) = 2 \times x_{1,j} + 1 \times x_{2,j} + \epsilon_j(t).$$

- $\epsilon_j(t)$ drawn independently for each arm in every trial, from $N(0, 1)$ in Experiment 1 and from $N(0, 0.25)$ in Experiment 2.
- Task was to maximize the cumulative reward.

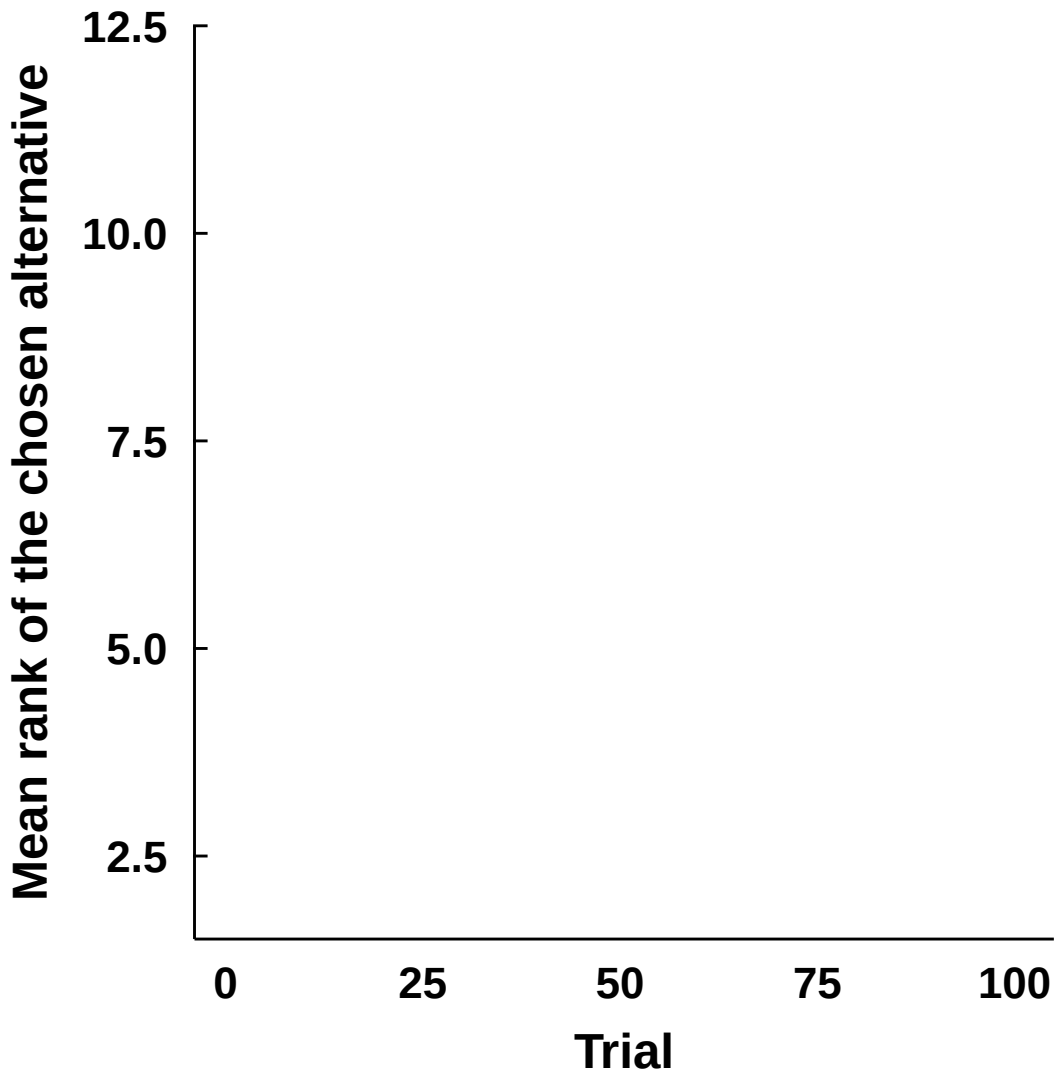
Screenshot - CMAB



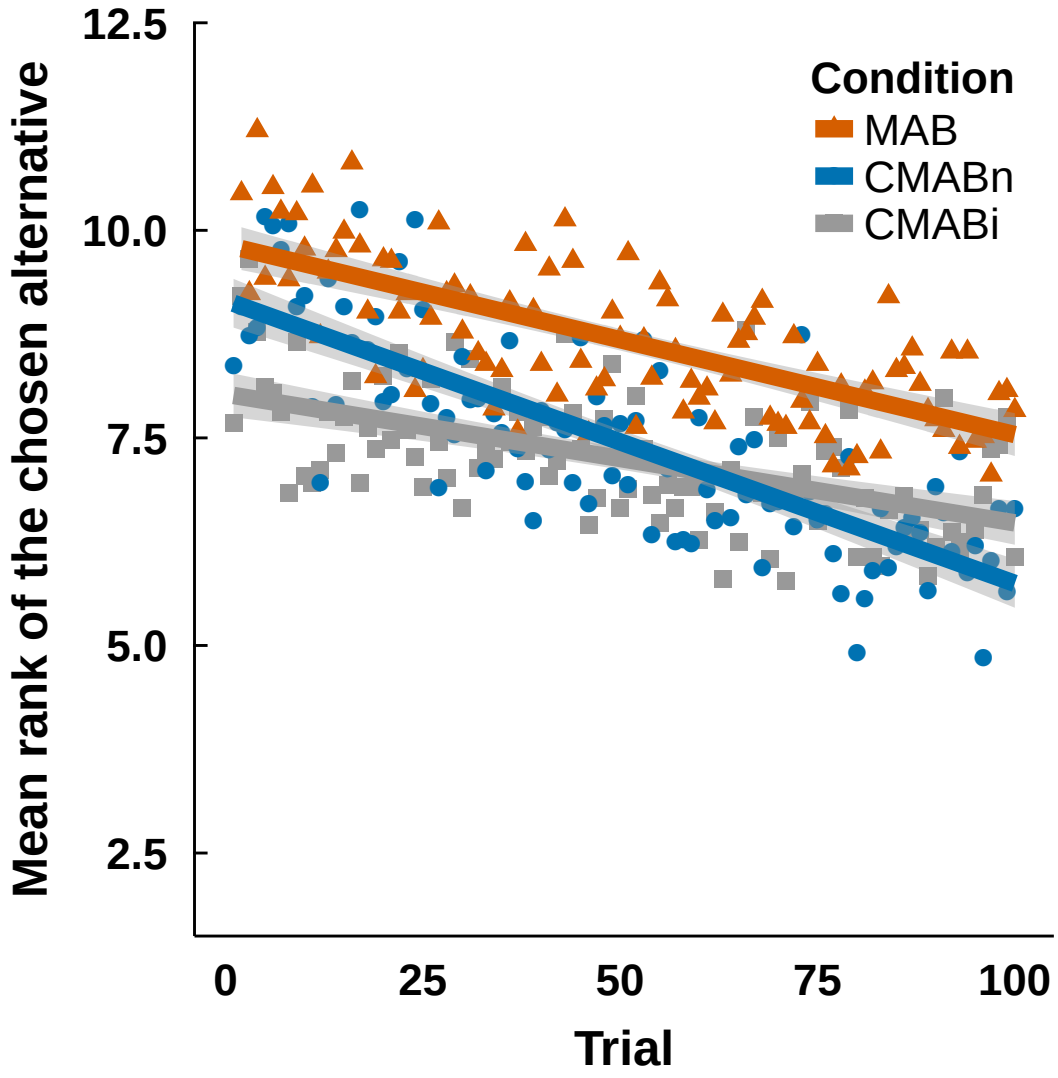
Screenshot - MAB



Behavior: Average choice rank

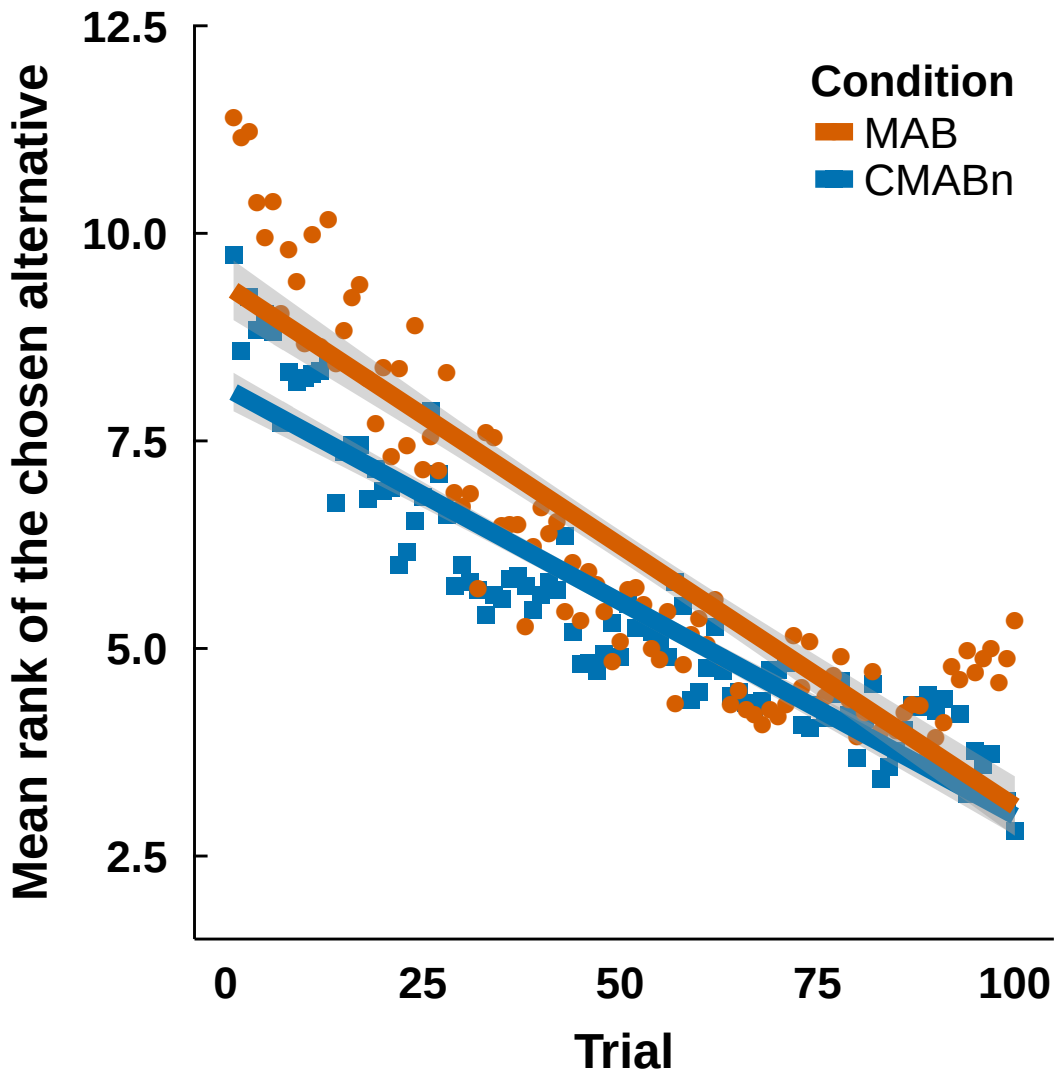


Behavior: Average choice rank - High noise



- Main effects significant, but interaction NOT.

Behavior: Average choice rank - Low noise



- Only block effect significant.

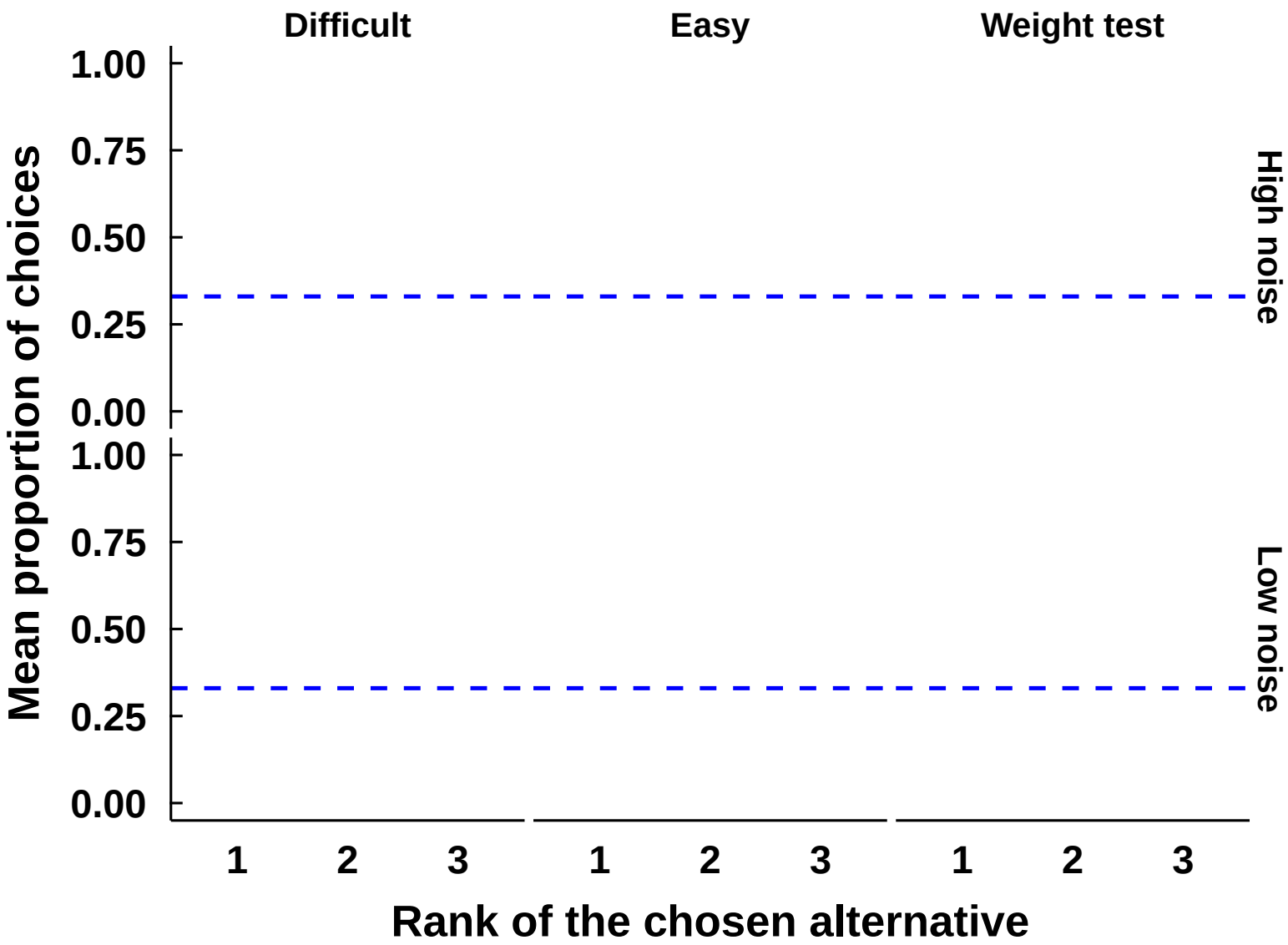
Behavior: One-shot choices in the test phase

Three alternatives:

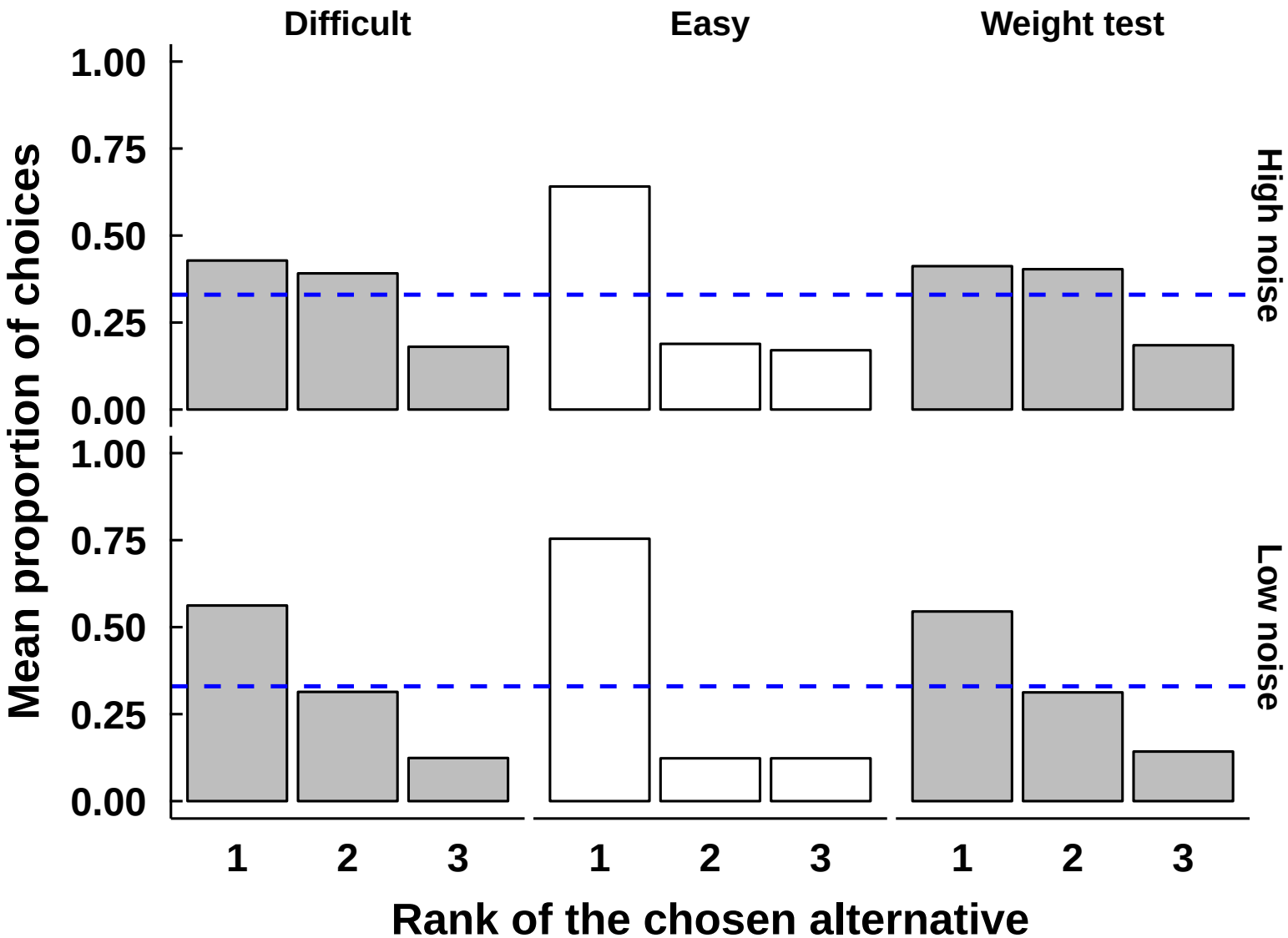
- **Dominating** - highest function value.
- **Neutral** - middle function value.
- **Dominated** - lowest function value.

One shot choices, 70 trials, no feedback!

Behavior: One-shot choices in the test phase



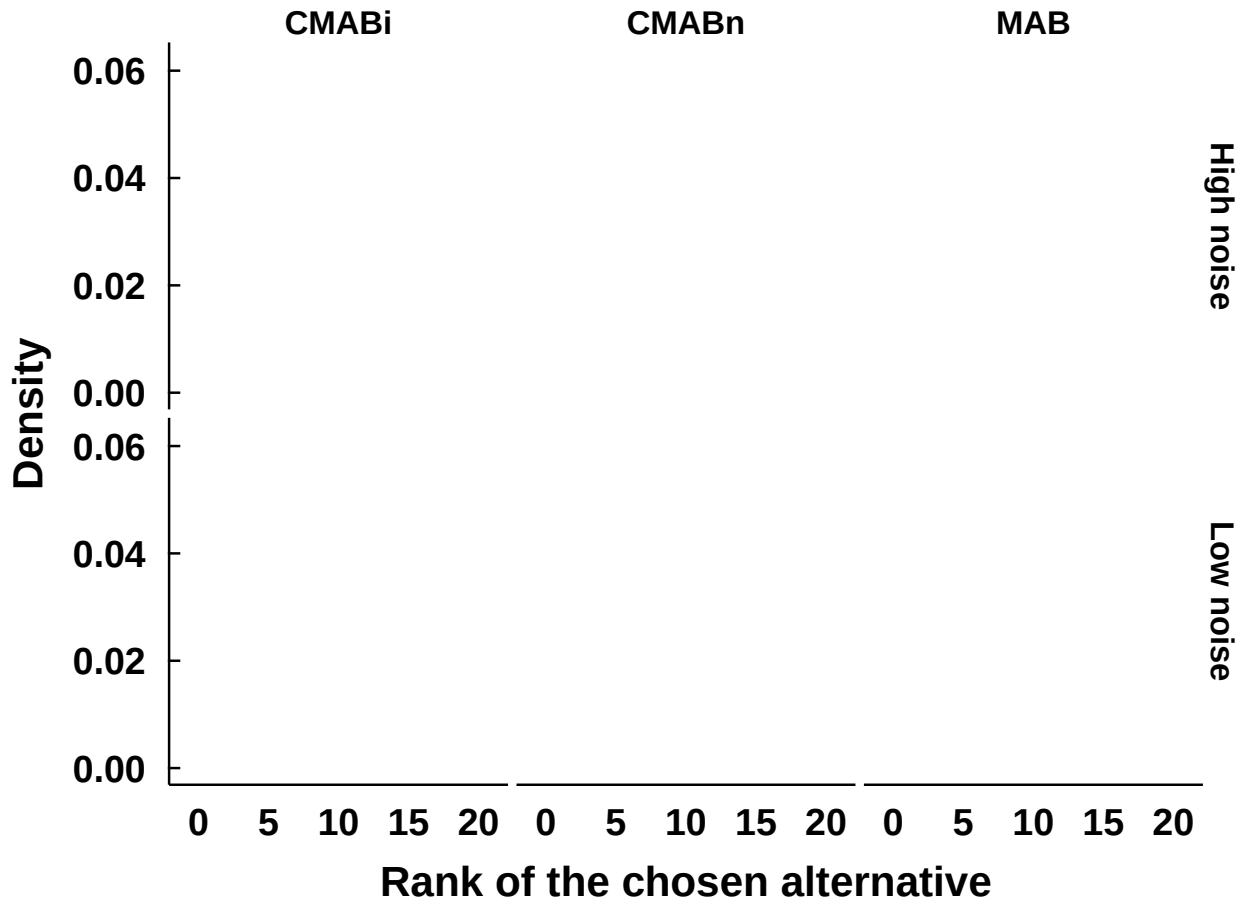
Behavior: One-shot choices in the test phase



Behavior: Ranks after switching in first 50 trials

Exploration guided by functional knowledge should be:

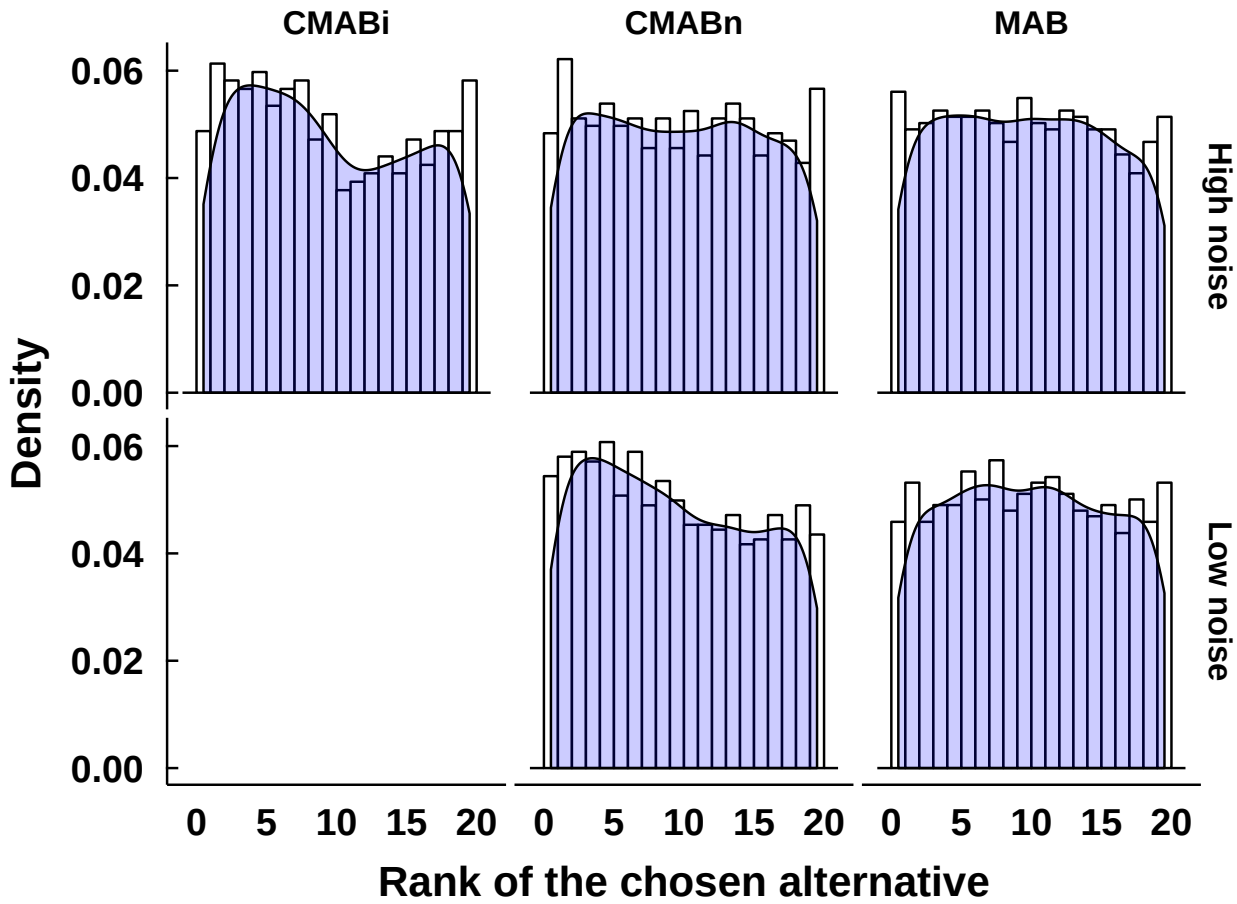
- Skewed toward highly ranked alternatives.
- Should include the extremes.



Behavior: Ranks after switching in first 50 trials

Exploration guided by functional knowledge should be:

- Skewed toward highly ranked alternatives.
- Should include the extremes.



Choice rule - Softmax

- All models used the softmax choice rule

$$P(C(t) = j) = \frac{\exp(\theta E_j(t))}{\sum_{k=1}^{K} \exp(\theta E_k(t))}$$

- θ is a parameter for sensitivity to expected value differences.

Naïve RL – Mean tracing

- **The delta rule** (Sutton & Barto, 1998):

$$E_j(t) = E_j(t-1) + \delta_j(t)\eta[R_j(t) - E_j(t-1)],$$

where $\delta_j(t)$ is an indicator variable (1 if alternative j was chosen on trial t, 0 otherwise) and fixed learning rate, $0 \leq \eta \leq 1$.

- **The decay rule** (Ahn et al, 2008) where expected values of the unchosen alternatives decay towards 0:

$$E_j(t) = \eta E_j(t-1) + \delta_j(t)R_j(t),$$

with decay parameter $0 \leq \eta \leq 1$.

Function learning based RL – LMS

- The **least-mean-squares (LMS)** network model (Gluck & Bower, 1987)

$$E_j(t) = \mathbf{x}_j \hat{\mathbf{w}}(t),$$

where $\hat{\mathbf{w}}(t)$ is a vector of estimated connection weights (identical for each alternative) and $\mathbf{x}$ is a feature vector.

- Updated through the delta rule

$$\hat{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) + \eta(R_j(t) - E_j(t))\mathbf{x}_j^T,$$

where η is a learning rate. Weights are initialized as $\hat{\mathbf{w}}(0) = (0,0)^T$.

Function learning based RL – LMS Decay/Delta

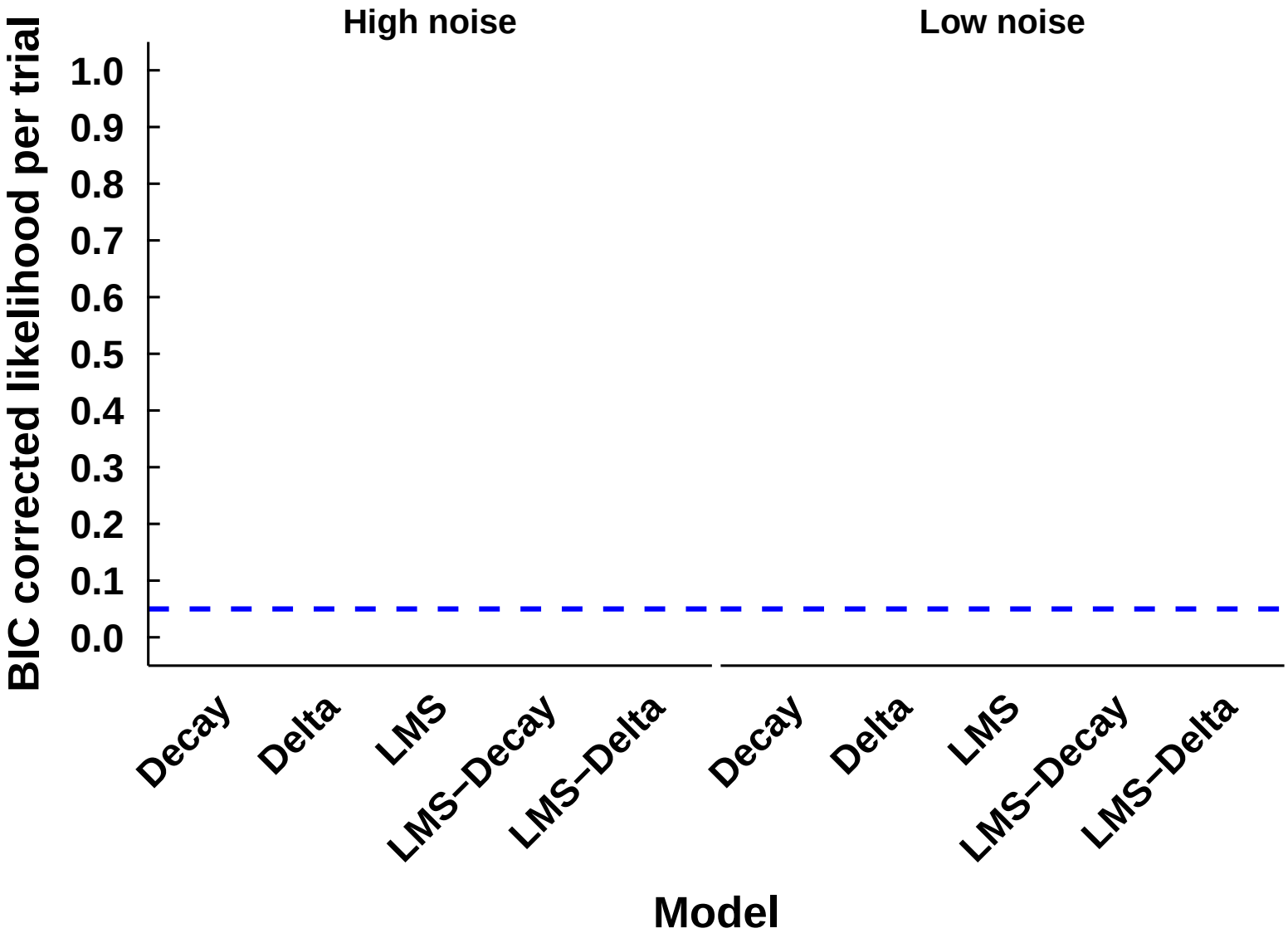
Expected values are weighted combination of:

1. expected values produced by the **LMS** network model with an addition of an intercept variable that learns average payoff, $E_j^{LMS}(t)$
2. expected values from either **Decay** or **Delta** mean tracing rule that learns average payoff of each alternative, $E_j^{Mean}(t)$

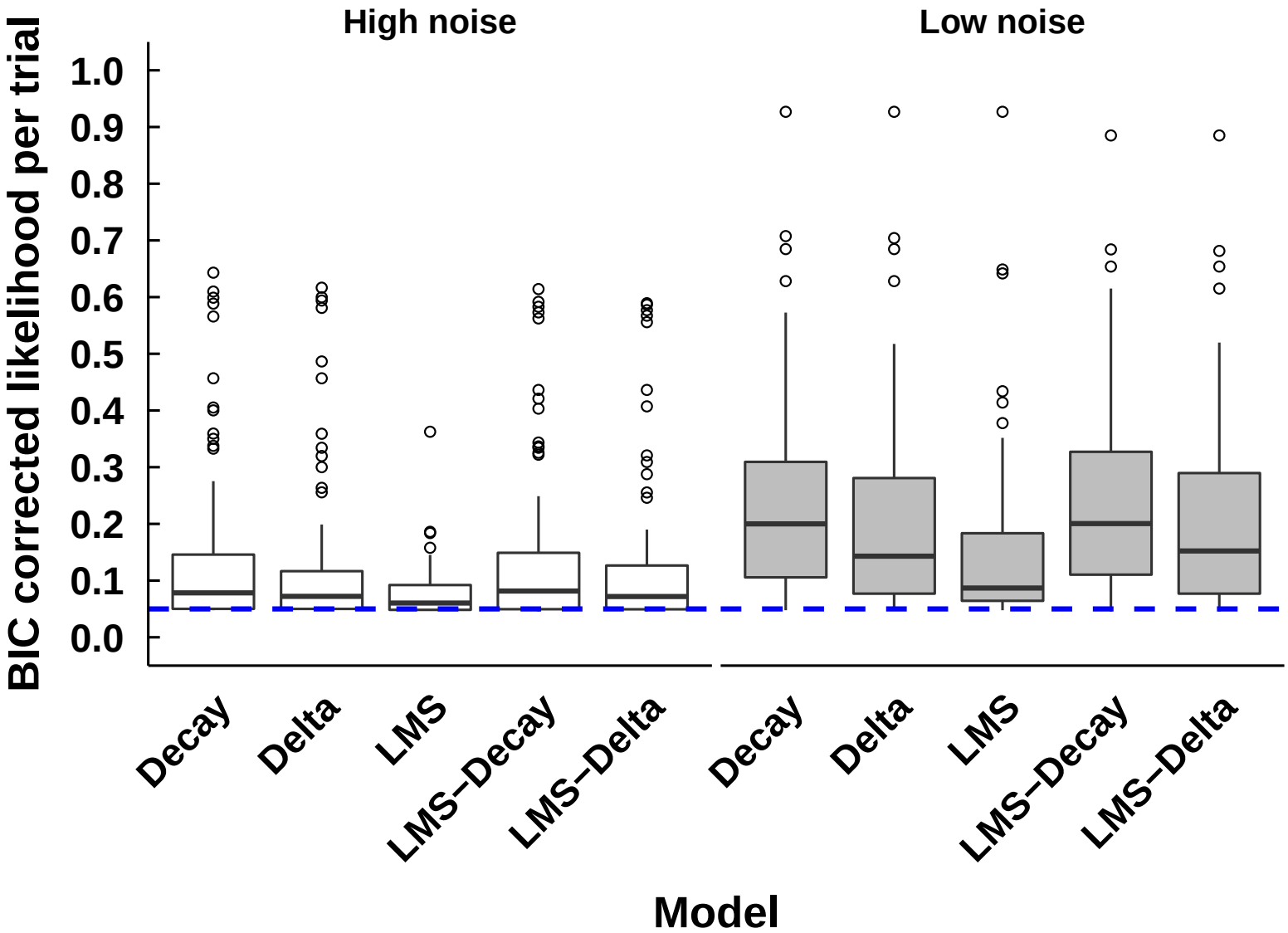
$$E_j(t) = \pi E_j^{LMS}(t) + (1 - \pi) E_j^{Mean}(t),$$

where π is a mixture parameter regulating how much person relies on function learning vs learning specific payoffs in making predictions.

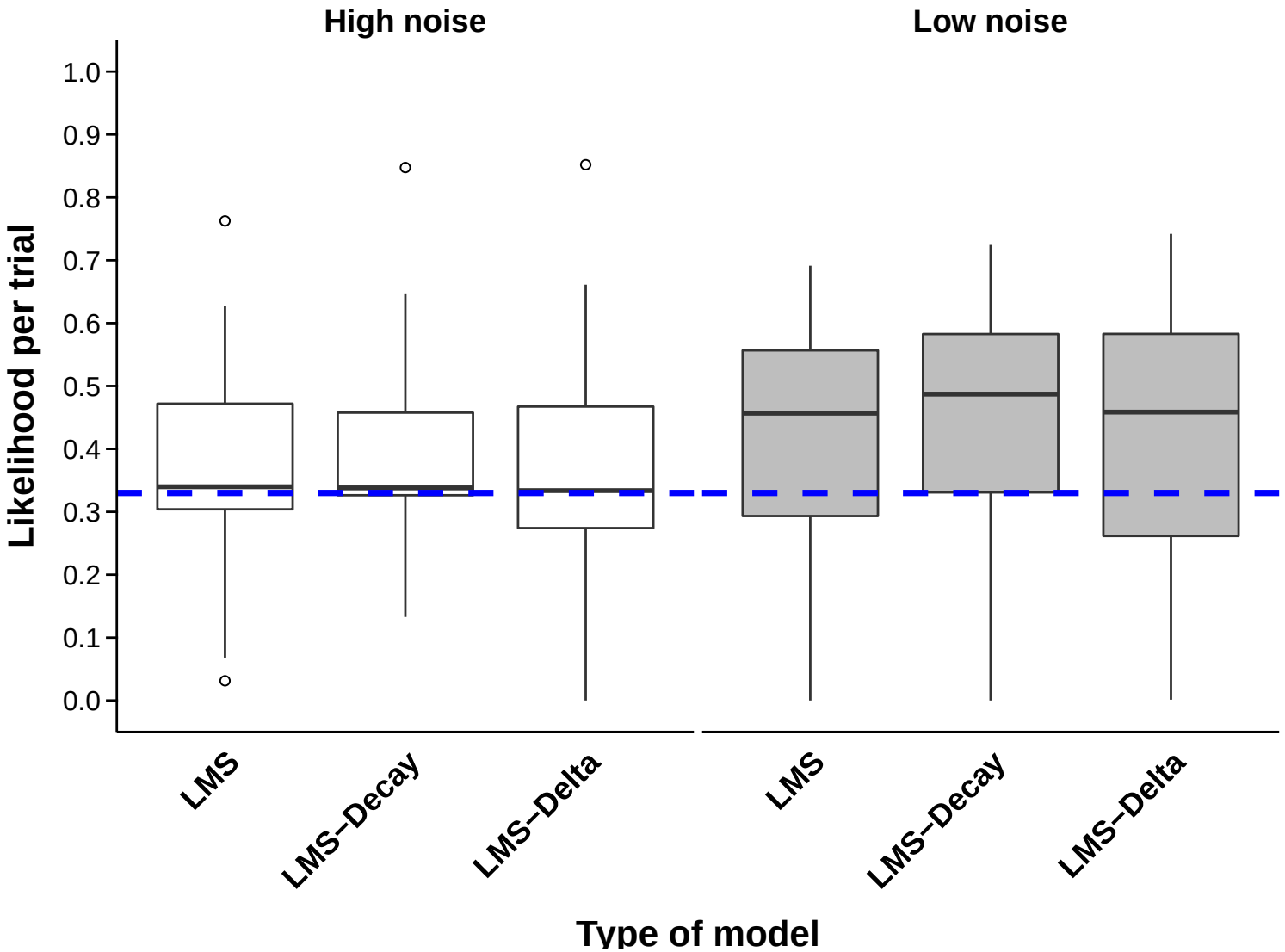
Model fits: Training phase – CMAB conditions



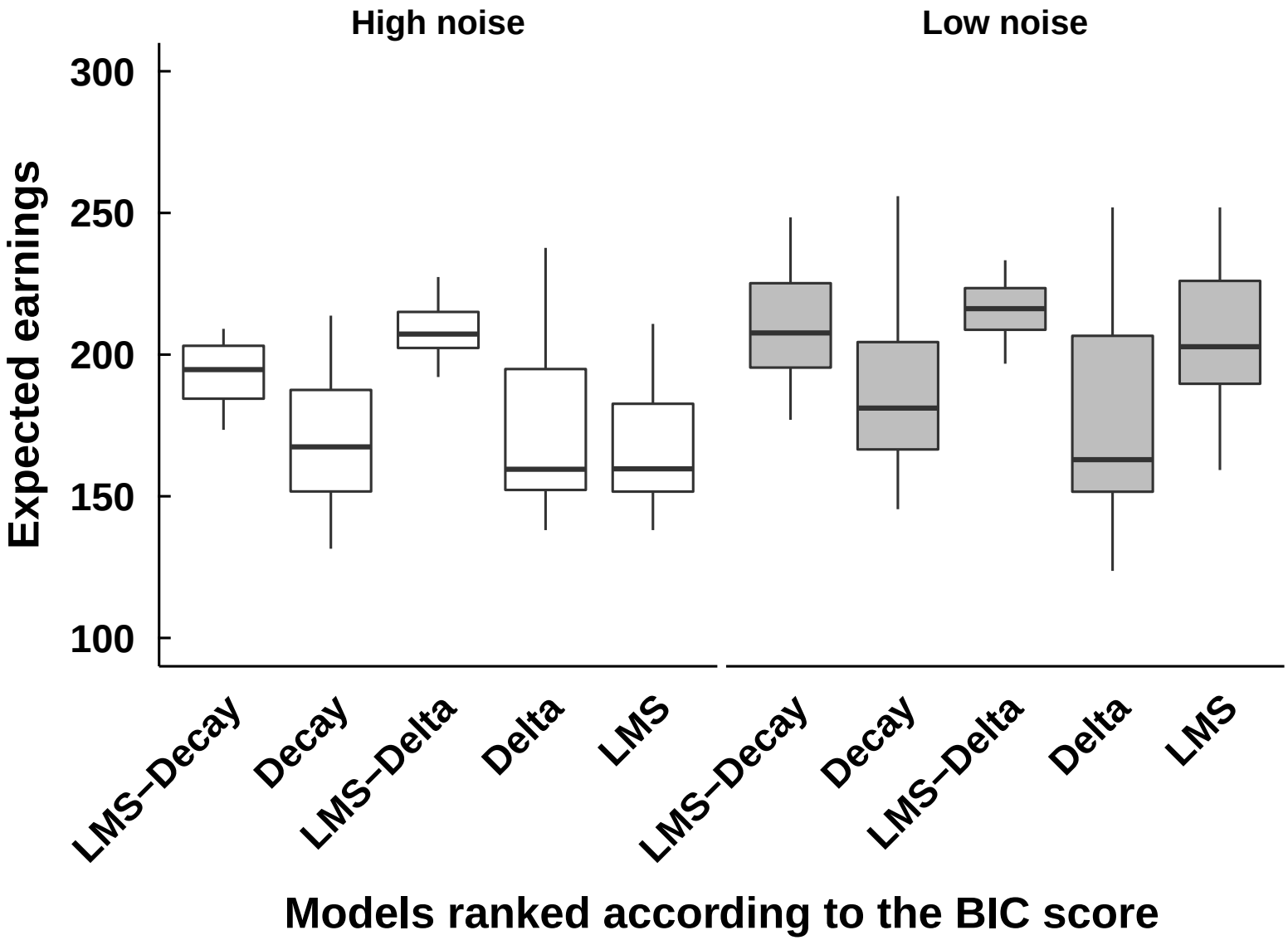
Model fits: Training phase – CMAB conditions



Model fits: Test phase - CMAB conditions



Training phase model fits and earnings – Different types of participants?



Summary

- CMAB - a promising new paradigm developed.
- Closer to the decisions in the wild.
- Initial modeling, but much room for improvement – Active learning!
- Distinguish classical and function based exploration – Need smarter experimental designs.
- Choice overload might be beneficial!
- Parallel modes or switching?

Acknowledgments

Hrvoje Stojic and Pantelis P. Analytis are on the job market!

Interested in working with us?

Let us know at:

hrvoje.stojic@upf.edu
analytis@mpib-berlin.mpg.de

Funding:

- FPU grant, Ministry of Education, Culture and Sports, Spain
- Barcelona Graduate School of Economics
- Max Planck Institute for Human Development, Berlin