Learning Dynamics of Semantic Knowledge

Jirko Rubruck[1], Andrew Saxe[2,3], Christopher Summerfield[1]

[1]Dept. of Experimental Psychology Oxford University, [2]Gatsby Computational Neuroscience Unit, UCL,
[3]CIFAR Azrieli Global Scholars program, CIFAR

Introduction

Human cognition relies on a rich set of semantic knowledge enabling and aiding reasoning about the world [1]. Much of this knowledge can be represented in hierarchical taxonomies. Simple artificial neural networks (ANNs) can extract such semantic structures via gradient descent [1, 2]. Saxe et al. (2019) extended these findings through careful analysis of semantic learning in deep linear networks [3]:

$$\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$$

The simplicity of the model allows for exact analytical solutions to learning dynamics [3]. A singular value decomposition of the time dependent weight matrices reveals the learning dynamics in the singular values $\mathbf{A}(t)$

$$\mathbf{W}^2(t)\mathbf{W}^1(t) = \mathbf{U}\mathbf{A}(t)\mathbf{V}^T$$

We employ key phenomena in these learning dynamics as predictions for our human behavioural task:

- ❖ Progressive differentiation and stage-like transitions

Methods

Participants learn about a synthetic dataset consisting of planet classes and associated properties in a binary tree. Properties are learned under the guise of a cover story of a space botanist that encounters new planets and must learn which plants grow on them.

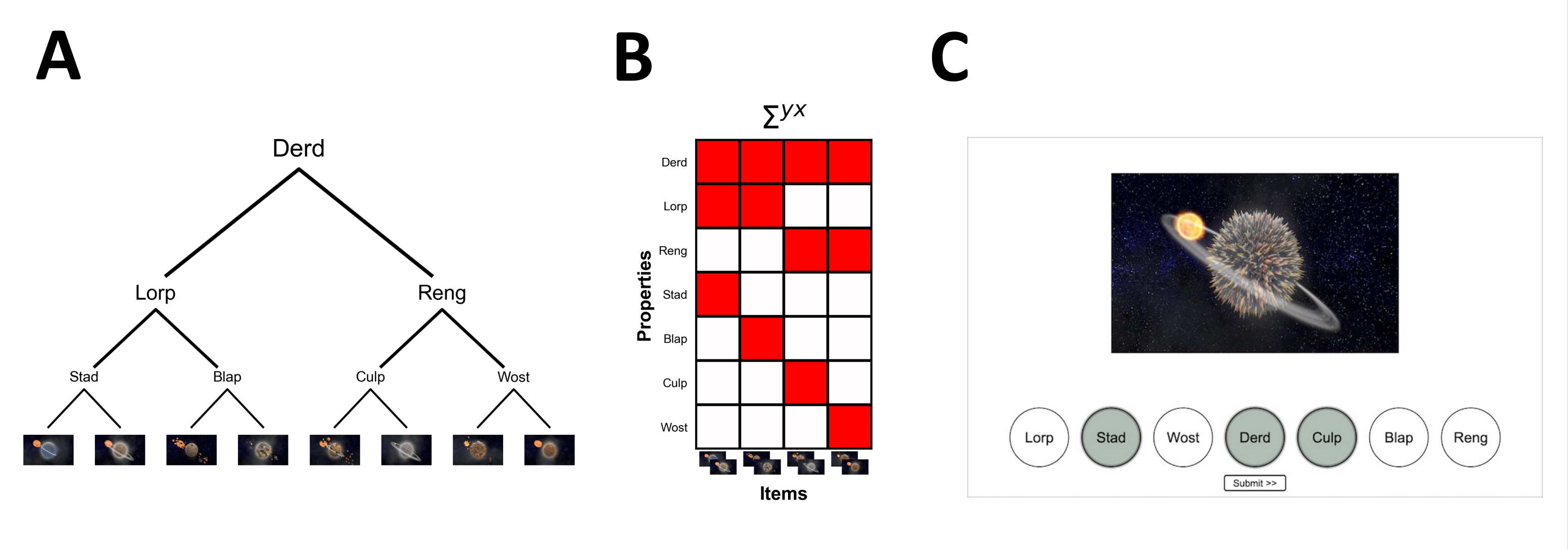

Figure 1. Experimental Design

On each trial participants are required to predict semantic properties of stimuli classes. *N* = 49, 8 blocks, 16 trials. Forced correction of each trial.

Results

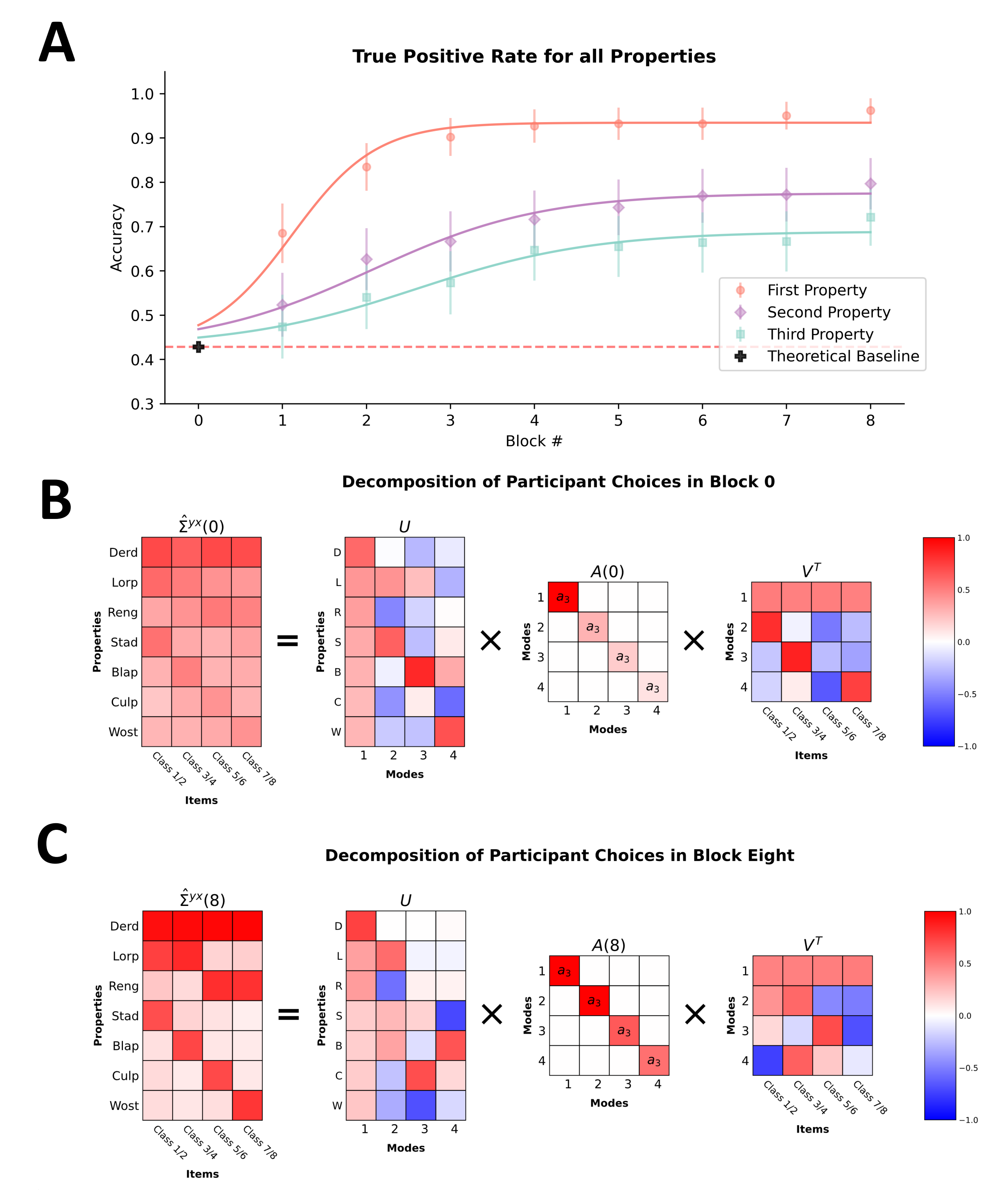

Figure 2. Choices and decompositions with respect to participant input-output correlation matrices. Learning respects, the stimulus hierarchy.

Results (cont.)

We then compare human and network input-output correlation matrices across different initializations.

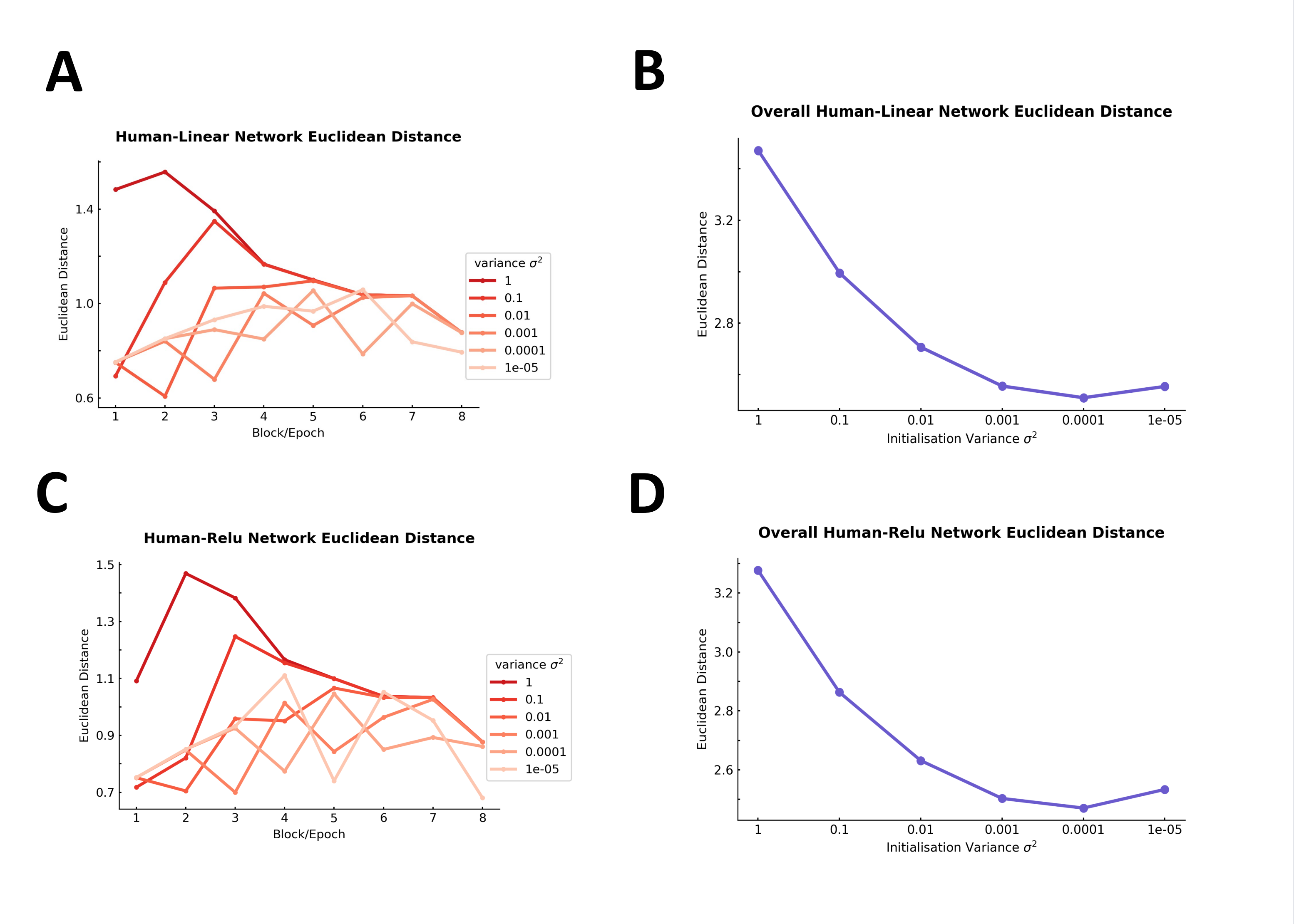

Figure 3. Euclidean distance between vectorised input-output-correlation matrices of humans and neural network input- output correlation matrices across different initialisations.

Future Directions

We conducted a longitudinal follow-up in which participants learn an extended hierarchy over the course of three days. We interleave the learning with there days of fMRI scanning to examine representational change throughout learning. We are currently analyzing this dataset.

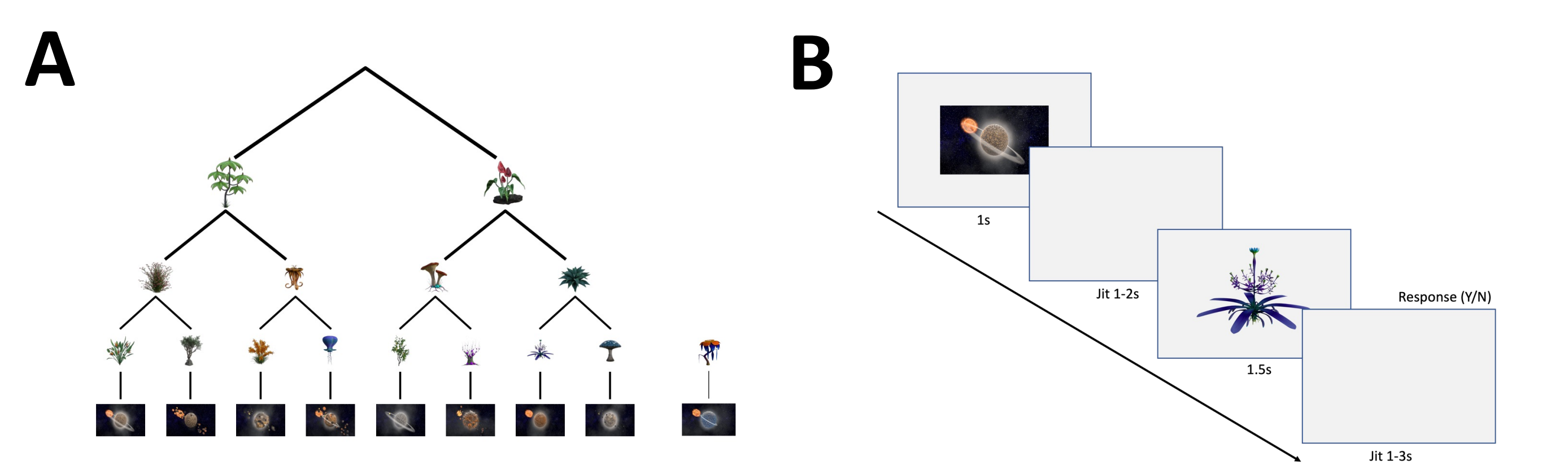

Figure 3. Follow-up fMRI experiment. A. Extended semantic tree with control planet. B. Scanning task.

Discussion

Participant learning respects the hierarchical structure of the data. Decomposition of participant choices emphasize this finding. The difference between the lowest and the mid-level of our hierarchy is remarkable as low-level properties are more specific identifiers of a particular class. Network simulations confirmed that human choices aligned more closely with networks that display stage-like transitions and progressive differentiation.

References

1. Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(6), 689.
2. Hinton, G. E. (1986, August). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society* (Vol. 1, p. 12).
3. Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537-11546.