Latent dimensions in neural representations predict choice context effects

Asaf Madar (asafmadar@mail.tau.ac.il)
Sagol School of Neuroscience, Tel Aviv University
Tel Aviv, Israel 69978

Tom Zemer (tomzemer@mail.tau.ac.il)
Sagol School of Neuroscience, Tel Aviv University
Tel Aviv, Israel 69978

Ido Tavor* (idotavor@tauex.tau.ac.il)
Faculty of Medical & Health Sciences, Tel Aviv University
Sagol School of Neuroscience, Tel Aviv University
Tel Aviv, Israel 69978

Dino J Levy* (dinolevy@tauex.tau.ac.il)
Coller School of Management, Tel Aviv University
Sagol School of Neuroscience, Tel Aviv University
Tel Aviv, Israel 69978

***** These authors contributed equally to this work.

Abstract:

Human choice is often affected by the context of available alternatives, a phenomenon known as choice context effects. To explain context effects, current models require the choice options to be described by two numerical attributes. However, decision-makers are not restricted by these attributes and might represent the options by additional latent attributes. Here, we propose using participants' neural representations to gain access to the full attribute set they consider, while relaxing the assumptions regarding their attribute space. We aimed to use these representations to predict the context effects in participants' choices. We estimated the context effects elicited by lottery stimuli using one behavioral sample ($n = 122$) and then recruited two independent fMRI samples in a preregistered design ($n_{first} = 28, n_{replication} = 34$) to estimate the neural representations of each lottery *without* the context of choice. We predicted the context effects based only on the neural similarity between the individual lotteries, improving out-of-sample predictions by 14% and explained variance by 20% compared to traditional methods. This framework can be generalized to any stimulus type and help extend the study of context effects to more naturalistic stimuli.

Keywords: Decision-making; Representational similarity Predictive modeling

Main

Human choices are known to be influenced by the interactions between available choice options and their attributes, a phenomenon known as choice context effects (Payne, 1982). One of the most well-known context effects in decision-making is called the *decoy effect* (Huber et al., 1982). It occurs when adding a third inferior "decoy" option to a set of two options, and while this third option is rarely selected, it influences the propensity of choosing one of the original options ("target") over the other ("competitor"). For example, one could add a medium-sized high-priced cup of coffee to a menu already including large high-priced and small low-priced cups, to increase the sales of the large cup instead of the small one. The effect has been studied extensively over the past four decades and explained by various computational models (Dumbalska et al., 2020; Herne, 1999; Simonson, 1989; Tsetsos et al., 2010; Usher et al., 2019).

When trying to explain choice context effects, researchers usually face two interacted problems. First, they try to understand the interactions between the available choice options, their attributes, and the effect on participants' choices. Second, researchers cannot describe the full attribute space of each multi-attribute choice option, due to multiple latent attributes that cannot be explicitly described. For example, a cup of coffee could be described only by its volume size and price, but also has sensory attributes such as taste, temperature, and color. The inaccessibility of these latent attributes usually leads researchers to describe the choice options using only explicit and numeric attributes in a two-dimensional attribute space such as price and quality (Fig 1a).

Here, instead of relying on the researchers' traditional two-dimensional view of the choice options to predict the decoy effects, we propose to rely on the decision-maker's view. To do so, we seek to estimate the choice options' representations as they are represented in the human brain in a data-driven way, with fewer assumptions regarding the structure of the underlying attribute space. In doing so, we provide a method which could be generalized to any higher-dimensional choice option by incorporating concepts from object representations and representational geometry into models of decision-making (Edelman, 1998; Kriegeskorte et al., 2008; Shepard & Chipman, 1970).

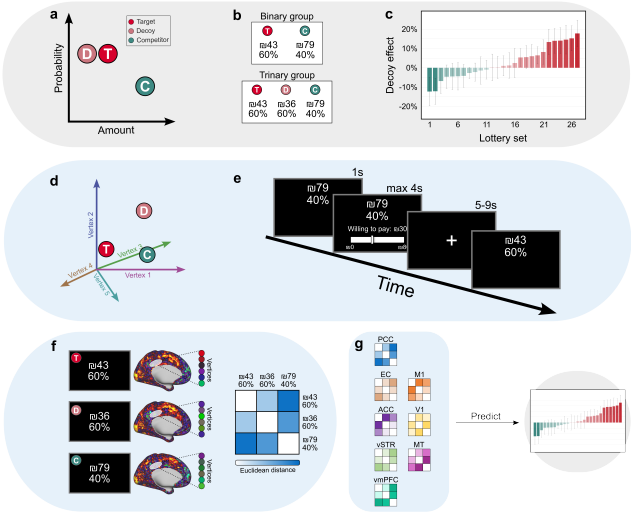

Figure 1. (a-c) Behavioral sample experiment. (d-g) fMRI samples experiment. See details in text.

Estimating the decoy context effects

First, we aimed to estimate the decoy effects elicited by a wide range of lottery stimuli. Each lottery is described by a probability of winning an amount of money and otherwise winning nothing (e.g., 40% to win $20, 60% to win $0). The probabilities ranged from 25%-75% and the amounts ranged from 4-79 NIS (≈$1-$20).

We recruited 122 participants who performed a standard decoy task (Fig 1b). They were randomly assigned to one of two groups (*binary* or *trinary*) and were asked in each trial to choose the lottery they prefer. In the *binary group*, participants were presented with two lotteries in each trial, the *target* and *competitor*. In the *trinary group*, participants were presented with three lotteries in each trial, which included the same two

lotteries as in the binary group and an additional *decoy* option. Participants were presented with 27 unique trials having either two or three lottery options.

The decoy effects were calculated as the change in the propensity to choose the target option in the trinary group and in the binary group. As can be seen in Fig 1c, the magnitude of decoy effects in our behavioral sample ranged from -12.3% to 17.8%. Ultimately, our main aim was to predict this variability using only the neural representations of the individual lottery stimuli.

The decoy effect is predicted by similarity of neural representations

To do so, instead of analyzing each lottery based on its location in the explicit two-dimensional attribute space of amount and probability (Fig 1a), we analyzed it based on its location in the high-dimensional space of neural representations (Fig 1d). We recruited two additional independent fMRI samples in a preregistered design ($n_{first} = 28$, $n_{replication} = 34$). Inside the MRI scanner, participants were presented with only one lottery on each trial for a total of 31 unique lotteries and were asked to state how much they are willing to pay in order to participate in that lottery (Fig 1e). The lotteries were identical to the ones presented to the behavioral sample. Participants completed five identical blocks of 31 trials each.

To predict the decoy effects using neural representations, we calculated the representational dissimilarity matrix (RDM; Fig 1f) for each participant in each of eight pre-defined regions, including value-related areas: Posterior and anterior cingulate cortices (PCC, ACC), ventromedial prefrontal cortex (vmPFC), ventral striatum (vSTR), and entorhinal cortex (EC), and sensorimotor areas: V1, M1, and middle temporal visual area (MT). We then trained and evaluated Lasso regression models to predict the decoy effects estimated from the behavioral sample using the eight RDMs averaged across the fMRI participants (Fig 1g).

Using the first fMRI sample, we significantly predicted the decoy effects using only the neural similarity between the lotteries with an average error of 6.6% between the predicted and actual decoy effects ($RMSE = 0.066, p = 0.0015$, Fig 2a; $r = 0.473$, Fig 2b). We replicated this result in the replication fMRI sample showing again high prediction success with an average error of 6.6% ($RMSE = 0.066, p = 0.0016$, Fig 2a; $r = 0.515$, Fig 2b). Importantly, these models did not have direct access to the explicit attributes of the lottery options and used only the neural representations of the pre-defined visual, motor, and value-related ROIs. To serve as a baseline of prediction performance, we also trained a regression model to predict the decoy effect using the explicit attributes of each lottery (amount and probability) in each lottery triad, which performed worse than the RDM models with an average error of 7.6% ($RMSE = 0.0764, p = 0.0109$, Fig 2a; $r = 0.314$, Fig 2b). This means our RDM models improved the predictions by 14% compared to the explicit attribute model.

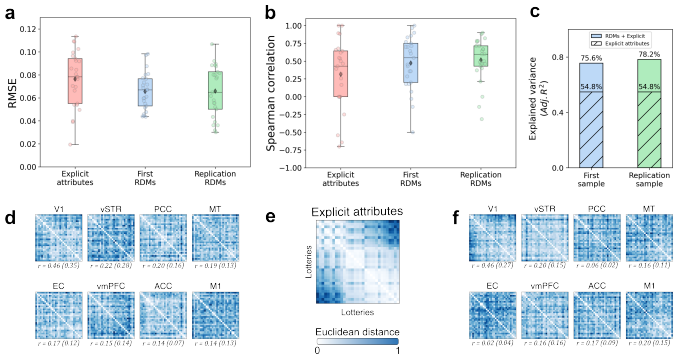

Figure 2. (a-b) Out-of-sample predictions. Each dot represents one fold. (c) Data fitting procedure. (d-f) Explicit attribute representation analysis. Details in text.

To test whether the variance explained by the RDM models extended beyond the explicit attributes, we also performed data fitting with combined regression models, which included both the explicit attributes and the RDMs of the selected ROIs (Fig 2c). The combined models explained 20% more variance compared to the model using only explicit attributes (Baseline: $R^2_{adj} = 0.548$; First: $R^2_{adj} = 0.756$; Replication: $R^2_{adj} = 0.782$). This shows that the neural representations enclose unique information that goes beyond the explicit attributes of amount and probability.

Lastly, we investigated whether the representations in our pre-defined ROIs primarily represented the lotteries' explicit attributes. To do so, we calculated the explicit attributes RDM for all lotteries in the two-dimensional attribute space (Fig 2e) based on the Euclidean distance between the amounts and probabilities of each pair of lotteries (Fig 1a). Then, we correlated the lower triangle of this attribute-RDM with the lower triangle of the neural RDM calculated for each ROI (Fig 2d, f). All pre-defined ROIs, except V1, had low correlations (ranging from $r = 0.019$ to $r = 0.221$), while V1 had a high correlation ($r_{first} = 0.456$, $r_{replication} = 0.457$) suggesting that V1 largely represented the explicit attributes while the other ROIs mainly represented latent attributes. We also repeated this analysis after controlling the perceptual similarity between the stimuli and found similar results (correlations shown in brackets in Fig 2d, f). Importantly, as our models used all pre-defined ROIs, they relied on a mixture of both latent and explicit attributes. Incorporating both types of attributes contributed to the superior performance of our models compared to traditional methods.

In this work, we showed the advantages of using high-dimensional neural representations to predict choice context effects.

Acknowledgments

The authors acknowledge with thanks the support of the Israel Science Foundation (1432/23).

References

Dumbalska, T., Li, V., Tsetsos, K., & Summerfield, C. (2020). A map of decoy influence in human multialternative choice. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(40), 25169–25178. https://doi.org/10.1073/PNAS.2005058117/-/DCSUPPLEMENTAL

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467. https://doi.org/10.1017/S0140525X98001253

Herne, K. (1999). The Effects of Decoy Gambles on Individual Choice. *Experimental Economics 1999 2:1*, *2*(1), 31–40. https://doi.org/10.1023/A:1009925731240

Huber, J., Payne, J. W., & Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research*, *9*(1), 90–98. https://doi.org/10.1086/208899

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/J.NEURON.2008.10.043

Payne, J. W. (1982). Contingent decision behavior. *Psychological Bulletin*, *92*(2), 382–402. https://doi.org/10.1037/0033-2909.92.2.382

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*(1), 1–17. https://doi.org/10.1016/0010-0285(70)90002-2

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, *16*(2), 158–174. https://doi.org/10.1086/209205

Tsetsos, K., Usher, M., & Chater, N. (2010). Preference Reversal in Multiattribute Choice. *Psychological Review*, *117*(4), 1275–1291. https://doi.org/10.1037/A0020580

Usher, M., Tsetsos, K., Glickman, M., & Chater, N. (2019). Selective Integration: An Attentional Theory of Choice Biases and Adaptive Choice: *Https://Doi.Org/10.1177/0963721419862277*, *28*(6), 552–559. https://doi.org/10.1177/0963721419862277