Identifying the neural dynamics of category decisions with computational model-based fMRI

Juliana D. Adema (juliana.adema@mail.utoronto.ca)*
Emily M. Heffernan (emily.heffernan@mail.utoronto.ca)*
Michael L. Mack (michael.mack@utoronto.ca)
Department of Psychology, University of Toronto
100 St George St, Toronto, ON M5S 3G3 Canada
* equal contribution

Abstract

Successful categorization requires a careful coordination of attention, representation, and decision making. Comprehensive theories that span levels of analysis are key to understanding the computational and neural dynamics of categorization. Here, we build on recent work linking neural representations of category learning to computational models to investigate how category decision making is driven by neural signals across the brain. We combine functional magnetic resonance imaging with hierarchical drift diffusion modelling to show that trial-by-trial fluctuations in neural activation from regions of occipital, cingulate, and lateral prefrontal cortices are linked to category decisions. Notably, lateral prefrontal cortex activation was also associated with exemplar-based model predictions of trial-by-trial category evidence. We propose that these brain regions underlie distinct functions that contribute to successful category learning.

Keywords: category learning, fMRI, computational modelling, EBRW, drift diffusion model

Introduction

Category learning, our ability to organize our experiences into meaningful concepts that can be leveraged in novel situations, is fundamental to the human experience. Not only are we able to group objects based on basic features such as colour and shape, but we are also capable of learning highly abstract and multivariate categories with relatively little practice. Besides commonplace categories such as "edible" and "friendly" and their antagonistic equivalents, objects can be assembled based on one or several complex perceptual features. Influential category learning models posit that novel objects are categorized according to their relative positions in a multidimensional psychological space populated with known category members in which the distance between objects determines their degree of similarity (Shepard, 1957). While this perceptual space can be composed of an unlimited number of dimensions, it is often the case that only some inform categorization decisions, and the weights of those few dimensions can vary (Nosofsky, 1986; Seger & Miller, 2010).

Although category learning has been studied for decades (Shepard, 1957; Young & Householder, 1938), more recent work has converged on a comprehensive account that formalizes the computational and neural mechanisms underlying successful learning (Zeithamova et al., 2019). While previous accounts of category learning feuded over the nature of concept representations (i.e., exemplar versus prototype), the multifactorial nature of concept learning has been found to implicate myriad neural systems. For instance, while lateral prefrontal cortex (LPFC) and parietal areas recruit specific category exemplars (Mack et al., 2013), the hippocampus and ventromedial prefrontal cortex inform representations based on category prototypes (Bowman & Zeithamova, 2018). In addition to similarity-based comparisons of representations, successful learning requires higher-order inferential processes, such as when faced with novelty or uncertainty invoked by a given stimulus (Paniukov & Davis, 2018). A comprehensive account must be able to link each of these processes; how these multiple brain systems interact remains to be defined.

Much of this work has focused on the link between predictions of representations formed during learning and how task strategies can influence what is learned by combining sophisticated neural analyses with formal model predictions (Bowman & Zeithamova, 2018; Braunlich & Seger, 2016; Mack, Love, & Preston, 2016; Mack et al., 2013). There is, however, a missing component: the decision making process itself. Limited work has explored the neural mechanisms that govern how category knowledge, once learned, is applied to novel situations to yield measurable behavioural changes in decision making.

Investigating the neural processes of category decision making requires a computational theory of how such decisions unfold. The Exemplar-Based Random Walk (EBRW) model (Nosofsky & Palmeri, 1997) formalizes category decisions as an evidence accumulation process. Evidence in support of different categories is sampled over time through similarity-based retrieval of category exemplars motivated by the seminal Generalized Context Model (GCM) (Nosofsky, 1986). The key innovation of EBRW is its ability to predict both response probabilities and speed, thereby providing a comprehensive formal account of the behaviour underlying categorization decision making.

Here, we leverage EBRW in an exploratory approach to identifying neural processes linked to categorization decision making. Our approach significantly extends prior work that has interrogated brain function with categorization models (e.g., Bowman & Zeithamova, 2018; Davis, Goldwater, & Giron, 2017; Mack et al., 2013) by 1) targeting neural response that fully characterizes decision responses and speeds, and 2) focussing on participant-specific predictions

through hierarchical model analyses. The EBRW model provides a strong theoretical framework for interpreting potentially informative neural processes; however, there are no analytic approaches for applying EBRW in a hierarchical manner that captures individual differences underlying common neural substrates at the group level in category decision making. As such, we split the two primary elements of EBRW—that is, exemplar-based category representations that drive an accumulation of noisy decision evidence—into a two-stage analytic approach: First, we interrogate neural signals related to decision making with a hierarchical variant of the drift diffusion model (DDM; Ratcliff, 1978), a computational model that approximates EBRW's decision making mechanism. Second, we evaluate the correspondence between the identified brain measures and individually tailored predictions of the GCM, the model that EBRW's exemplar-based category representations are based on. Since EBRW formalizes that category representations impact decision making through changes to the rate of evidence accumulation (Nosofsky & Palmeri, 1997) and that this sort of category evidence has been correlated with neural function in prefrontal and parietal cortices and striatum (Davis et al., 2017), we focus on the drift rate parameter in the DDM.

Thus, we test the hypotheses that brain activation during a classic categorization task (Mack et al., 2013; Medin & Schaffer, 1978) corresponds to the rate of category evidence accumulation on a trial-by-trial basis and that such neural decision signals may vary when category knowledge is generalized to novel relative to previously encountered stimuli. Given the exploratory nature of this approach (Thompson, Wright, & Bissett, 2020), we take a purposefully uninformed view of brain function. Specifically, we interrogate neural response from a parcellation of distinct regions across the whole brain independently identified in a large-scale, data-driven analysis of resting state interregional connectivity (Schaefer et al., 2018). This approach provides a key first approximation of how to identify neural function underlying complex human behaviour through the lens of a formal computational theory.

Methods

The current study leverages a previously published open-access dataset (Mack et al., 2013). This dataset, which includes behavioural and structural and functional magnetic resonance imaging (fMRI) data, was downloaded from OSF (https://osf.io/62rgs/). For completeness, we include a full description of the methods for this experiment.

Participants

Twenty-three participants participated in the experiment. Two participants were removed prior to analysis for excessive head motion during fMRI scanning, and one participant was removed for failure to learn the categorization task. The remaining 20 participants were included in the primary analysis (age range of 19–33 years; mean age of 23.5 years; 14 female).

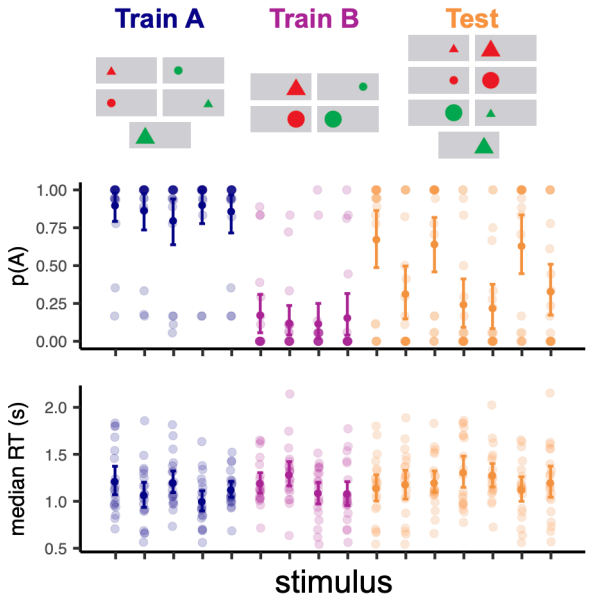

Figure 1: Stimuli and test performance. Stimuli composed of four binary dimensions were split into training (five A and four B items) and testing sets (example stimulus set shown in top row). Test performance showed typical accuracy (middle) and median reaction time (bottom) performance as in previous reports (e.g., Medin & Schaffer, 1978). Lighter dots depict participant-specific averages, darker dots depict group averages, and errors bars depict 95% confidence intervals.

Stimuli

The stimulus set was composed of 16 objects consisting of simple shapes enclosed in a grey, horizontally oriented rectangle (Figure 1). The simple shape varied based on four salient binary-valued features (colour: red or green, shape: circle or triangle, size: large or small, and position: right or left). For each participant, the four features were randomly assigned to the four dimensions defined by the 5/4 category structure (Medin & Schaffer, 1978). This structure is divided into two categories with the prototype member of category A corresponding to [0,0,0,0] and the prototype member of category B corresponding to [1,1,1,1]. Nine objects served as the training items with five for category A and four for category B. The remaining seven objects served as a test set.

Procedures

After an initial screening and consent in accordance with the University of Texas Institutional Review Board, participants were instructed on the category learning task. These instructions explained that the participant would be shown simple objects composed of different features and that the task was to learn which object belonged to one of two categories through corrective feedback.

Participants performed the training phase of the experiment in a behavioural testing room. On each training trial, one of

the nine training stimuli was displayed for 3.5s and participants made a response to the stimulus' category by pressing one of two labelled keys on the keyboard. Then, a fixation cross was presented for 0.5s, followed by a feedback display that presented the stimulus, the correct category, and whether the participant's response was correct or incorrect for 3.5s. Trials ended with a 0.5s fixation cross. The nine training stimuli were presented 20 times in randomized order during the initial training outside the scanner. Participants also completed additional training trials inside the fMRI scanner during an anatomical scan as a refresher of the training items' category membership. In total, across the entire training phase, participants completed 24 repetitions with each training stimulus.

After training, participants performed the testing phase during functional scanning. On each test trial, one of sixteen stimuli (consisting of the nine training stimuli and seven novel transfer stimuli) was displayed for 3.5s and participants made a category response by pressing one of two buttons on an MRI-compatible button box. A fixation cross was then presented for 6.5s. No feedback was provided during the testing phase. The 16 stimuli were presented three times in randomized order during six functional runs for 18 total repetitions per stimulus.

fMRI Data Acquisition and Preprocessing

Whole-brain imaging data were acquired on a 3.0T GE Signa MRI system (GE Medical Systems). Structural images were acquired using a T2-weighted flow-compensated spin-echo pulse sequence (TR=3s; TE=68ms, 256x256 matrix, 1x1mm inplane resolution) with thirty-three 3-mm thick oblique axial slices (0.6mm gap), approximately 20° off the AC-PC line. Functional images were acquired with an echo planar imaging sequence using the same slice prescription as the structural images (TR=2s, TE=30.5ms, flip angle=73°, 64x64 matrix, 3.75x3.75 in-plane resolution, bottom-up interleaved acquisition, 0.6mm gap). An additional high-resolution T1-weighted 3D SPGR structural volume (256x256x172 matrix, 1x1x1.3mm voxels) was acquired for registration and brain parcellation.

Anatomical and functional MRI data for each participant were preprocessed using the fMRIPrep automated MRI workflow (version 1.0.15; Esteban et al., 2019), which included brain extraction, motion correction, coregistration between functional and T1 volumes, and normalization to the MNI 2009c asymmetric brain template. AROMA-identified noise components were regressed out of the functional timeseries. For each run, trial-level beta parameters were estimated from the functional timeseries across the whole brain using the LS-S approach (Mumford, Turner, Ashby, & Poldrack, 2012). Finally, beta estimates across trials were averaged within 100 regions of interest (ROI) as defined by a resting-state brain parcellation (Schaefer et al., 2018) and an additional eight ROIs from subcortical regions including right and left hippocampus, caudate, putamen, and thalamus.

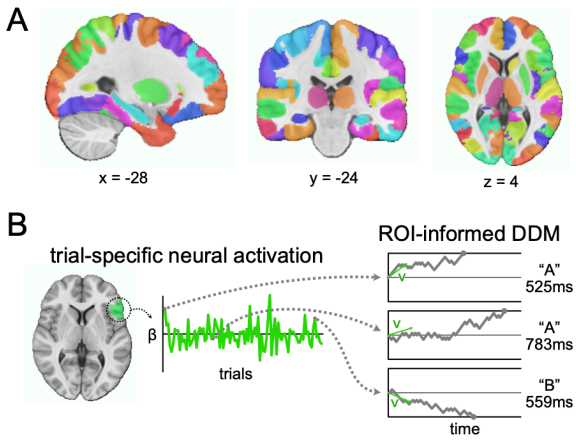

Figure 2: Brain parcellation and analysis schematic. A) Analyses targeted 108 brain ROIs derived from a data-driven resting-state functional parcellation and anatomically-defined subcortical regions. B) Mean trial-by-trial beta estimates within each ROI were extracted and leveraged to predict trial-by-trial changes in drift rate in DDM simulations of categorization decisions.

Brain-informed Drift Diffusion Modelling

DDM analyses were conducted by first averaging beta estimates within each ROI for each trial resulting in an ROI-specific timeseries (Figure 2B). Separate DDM simulations were conducted for each ROI wherein trial-by-trial changes in drift rate, v, were linked to the ROI timeseries. Given the theorized relationship in EBRW between exemplar-based category similarity and the random walk accumulation of evidence (Nosofsky & Palmeri, 1997), we focused on the link between brain and drift rate in the DDM. The DDM formalizes other parameters that distinctly influence predictions of decision making behaviour (e.g., decision threshold, α, and non-response time, T_{ER}); however, the key prediction from EBRW we test here depends on drift rate. Additionally, the relationship between drift rate and ROI activation was allowed to vary by stimulus type (training vs. testing items). Simulations were implemented with the Hierarchical Drift Diffusion Model (HDDM) library (Wiecki, Sofer, & Frank, 2013), which performs hierarchical Bayesian parameter estimation. MCMC sampling was conducted for 20,000 samples with 10,000 burn-in and thinning set to 2. The deviance information criterion (DIC) for each ROI-based model was compared to a baseline model not linked to neural activation. ROI-based models with smaller DIC values differing from the baseline model by at least 10 were consider a significantly better fit (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Parameter estimates from ROI-based models meeting this criterion were further explored by analyzing the posterior distributions of effects due to neural activation and stimulus type (training vs. testing items) on drift rate. Specifically, the probability of direction, *pd*, for each effect was calculated as the proportion of posterior samples in the most probable direction (i.e., *pd*

ranges from 0.5 to 1 with values closer to 1 for more likely effects).

Linking Category Evidence and Neural Response

Leveraging the HDDM provides a quantitative means for interrogating the link between neural activation and categorization decisions. In particular, finding that trial-by-trial fluctuations in the neural activation of certain brain regions predicts behaviour suggests these brain regions are performing an important role in mapping sensory information onto category knowledge. However, the DDM is agnostic to specific mechanisms underlying computations of category evidence, thus it represents only one component of the broader EBRW framework.

To test the hypothesis that the degree of evidence for one category over another is reflected in the neural dynamics of the identified ROIs, we performed an additional analysis that linked participant-specific cognitive model predictions of categorization to neural activation in the ROIs identified by the brain-informed DDM analysis. Specifically, we generated predictions of category evidence with the GCM (Nosofsky, 1986). A key mechanism of GCM is selective attention, whereby diagnostic feature dimensions for a given task are weighted to varying degrees in calculating similarity to stored category exemplars and thus modulate the evidence for each category. It has been previously shown that individual differences in both categorization performance and neural representations (e.g., Mack et al., 2013; Braunlich & Love, 2019) are related to attention weighting in the GCM.

Thus, to quantify participant-specific predictions of category evidence, we fit to each participant's behaviour by optimizing GCM parameters of dimensional attention weights (w) and sensitivity (c) with a genetic algorithm approach (differential evolution in *scipy* version 1.3.0) to maximize likelihood of response probabilities during the last block of training (Mack et al., 2013). Participant-specific optimized parameters were then leveraged to predict for each stimulus the degree of discriminatory evidence (ev) for the most probable category. This measure of category evidence for stimulus x is the absolute value of the difference between the summed similarity for category A and B:

$$ev_x = \left| \sum_{y \in A} e^{-c \sum_{i \in d} w |x_i - y_i|} - \sum_{y \in B} e^{-c \sum_{i \in d} w |x_i - y_i|} \right|$$

where d is the set of feature dimensions, A and B are the set of training items in the two categories, w is the set of optimized dimension weights, and c is sensitivity. We then evaluated the relationship between category evidence (ev) and trial-by-trial neural activation with a mixed effects linear regression. Specifically, category evidence was modelled as the response, neural activation as the predictor, and random intercepts were included for participants. Given that category evidence is exponentially distributed, a log link function was included in the regression model. Regression analyses were conducted with Bayesian estimation (*rstanarm* R package version 2.19.2). Neural activation models were compared to a baseline model that only included random intercepts.

Results

Testing phase performance (Figure 1) showed typical results consistent with previous reports (e.g., Medin & Schaffer, 1978). Responses to A and B training items demonstrated clear learning that was retained during test. Responses to novel test items varied according to match to category exemplars as determined by each participants' learning performance (Mack et al., 2013). Median reaction times (RTs) at the group level did not vary across items; however, there was variability across participants and trials. The brain-based DDM analysis offers a means for accounting for trial-by-trial variability in response choices and times with neural function.

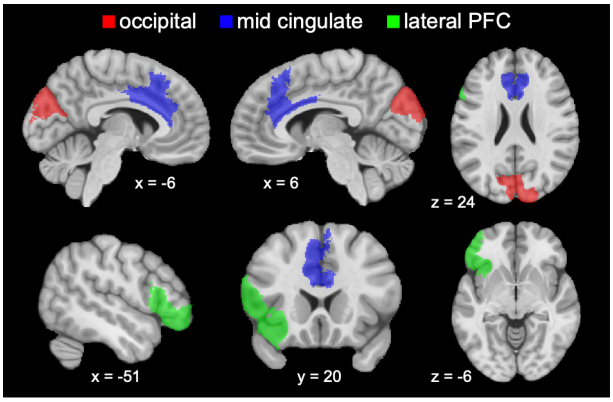

Figure 3: Brain regions linked to category decisions. DDM models informed by neural activation from occipital (red), mid cingulate (blue), and lateral PFC (green) each accounted for category decisions significantly better than the baseline model not informed by brain signals.

Brain Regions Related to Category Decisions

Across the brain, only six of the tested ROIs showed brain-informed HDDM predictions with significantly better accounts of category decisions relative to the baseline model. Interestingly, these six ROIs were composed of three pairs of adjacent regions, with each pair showing similar effects. To simplify presentation of the results and to best reflect the nature of their similar effects, we combined these pairs into three ROIs (Figure 3): an occipital region including extrastriate cortex, a region of midcingulate cortex, and a region of left lateral prefrontal cortex (PFC) extending into insula. HDDM simulations of these combined ROIs demonstrated significantly lower DICs (occipital: 6,712.9, mid cingulate: 6,713.5, lateral PFC: 6,711.5) relative to baseline (6,725.3).

Although these three regions demonstrated similar overall fits to behaviour, the nature of the relationship between ROI activation and drift rate was unique across ROIs (Figure 4). In the occipital ROI, neural activation was positively related to drift rate but only for test items ($pd_{train} = 0.8$, $pd_{test} = 0.998$).

In mid cingulate, neural activation was negatively related to drift rate but only for training items ($pd_{train} > 0.999$, $pd_{test} = 0.676$). Finally, the lateral PFC region showed a negative relationship between neural activation and drift rate for both stimulus types ($pd_{train} = 0.997$, $pd_{test} = 0.985$).

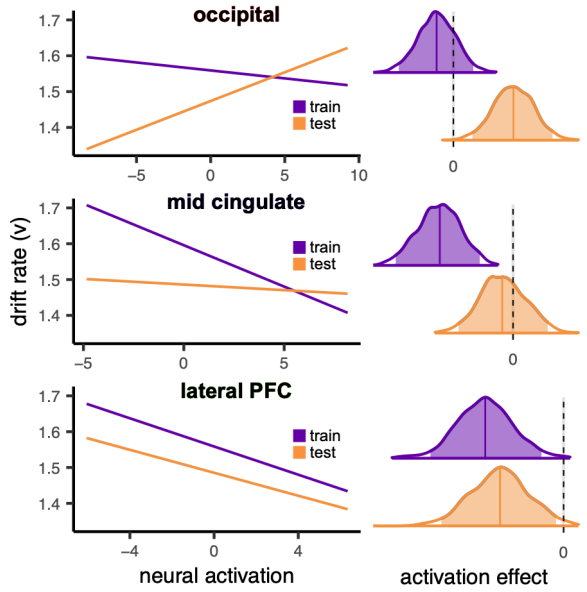

Figure 4: Effects of neural activation on drift rate. Posterior predictions of drift rate as a function of ROI activation (top: occipital, middle: mid cingulate, bottom: lateral PFC) separately for training (purple) and test (orange) stimuli are shown on the left. Posterior distributions of the activation effects on drift rate for training and test stimuli relative to 0 (dotted line) are shown on the right. Shaded regions represent 95% prediction intervals.

Category Evidence in Neural Activation

The key prediction of EBRW follows that category decision making is driven by similarity-based comparisons to category exemplars. Extending this hypothesis to neural function, it follows that brain regions key for category decision making will exhibit activation profiles that track category evidence. To test this prediction, we evaluated the association between category evidence, as derived from GCM-based model fits of learning behaviour, and neural activation in the DDM-identified brain regions.

Of the three ROIs, only lateral PFC showed a significant relationship with model-based predictions of category evidence (Figure 5; $R^2 = 0.145$) with higher lateral PFC activation associated with less discriminatory category evidence ($\beta = -0.004$, $CI = [-0.007, -0.001]$, $pd = 0.987$). These findings support the hypothesis that lateral PFC activation fluctuates as a function of category evidence (Paniukov & Davis, 2018) and that this category evidence plays a role in the accumulation of evidence in category decisions (Nosofsky & Palmeri, 1997).

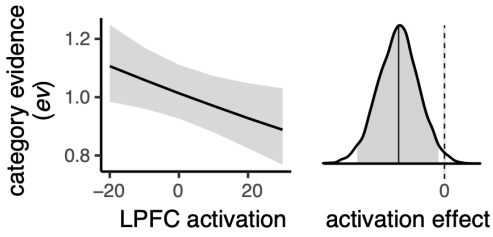

Figure 5: Relationship between lateral PFC activation and GCM-predicted category evidence. Shaded ribbon and regions depict 95% prediction intervals.

Discussion

By integrating a formal category decision making model, EBRW, with whole-brain neural measures, we demonstrate that activation in specific brain regions relates to the trial-by-trial dynamics of category decisions. Specifically, we found that activation in LPFC was associated with category decisions and the speed of those decisions. Notably, trial-by-trial activation in this region also related to exemplar-based predictions of category evidence. In both cases, the link to LPFC activation was an inverse relationship: higher activation was accompanied by slower decisions and less category evidence. These findings support an account of LPFC engagement tied to the difficulty of the current category decision, such that LPFC is recruited to resolve conflicts and help drive decision making in ambiguous circumstances.

The LPFC has been previously reported to be involved in a variety of tasks consistent with this interpretation. Monkey studies suggest that neurons in LPFC are sensitive to multidimensional feature representations (Mendoza-Halliday & Martinez-Trujillo, 2017) and code for category boundaries (Seger & Miller, 2010). Additionally, human work points to a specific role for LPFC in the control of memory retrieval, particularly in the face of ambiguity and competing information (e.g., Badre & Nee, 2018; Thompson-Schill et al., 1998). These prior results portray the LPFC as the recipient of perceptual information, to which it applies representations of category structure to guide and drive category decisions. When the perceptual representations provide less obvious category distinctions, an increase in activation may be indicative of a greater amount of cognitive effort required to make a decision.

Recent work has isolated the distinct contribution of subregions of LPFC, namely rostrolateral PFC (RLPFC), to categorization. It has been proposed that RLPFC is involved in tracking higher-order relations between object features, while other LPFC regions drive the decision making process more generally (Davis et al., 2017). For instance, a recent study of category decision making has determined that the RLPFC engages in comparisons of *dissimilarity* between exemplars, signalling its involvement in highly abstract category learning processes, while DLPFC is activated in cases of uncertainty when there is less similarity-based

evidence available to inform a category decision (O'Bryan, Worthy, Livesey, & Davis, 2018). Furthermore, RLPFC has been shown to continually evaluate categorization rules, even if the correct strategy has been arrived at (Paniukov & Davis, 2018). Therefore, when less information is provided by perceptual representations, or when the categorization decision is not obvious enough to rely on earlier cortical areas, LPFC is engaged to compare the current stimuli to previous exemplars according to task-specific goals. By directly linking trial-by-trial neural signals from LPFC to category decision making and the expression of exemplar-based category knowledge, our findings uniquely support the notion that LPFC is tracking category evidence in a behaviourally-relevant manner (O'Bryan, Walden, Serra, & Davis, 2018) akin to the mechanisms of EBRW (Nosofsky & Palmeri, 1997).

Although occipital and mid cingulate regions were not associated with model-based predictions of category evidence, activation in these regions did exhibit distinct relationships with evidence accumulation as formalized in the DDM. In occipital cortex, decision-related activation was restricted to novel test items, such that higher activation was accompanied by quicker and more accurate responses. Prior studies support the notion of concept representation in perceptual cortices. Specifically, it is thought that recurrent feedforward/feedback loops with medial temporal lobe and PFC allow visual regions to make inferences about stimulus features (Hindy, Ng, & Turk-Browne, 2016; Lee & Mumford, 2003). Moreover, occipital regions respond to the similarity between particular stimuli and category representations (Braunlich & Love, 2019), as well as demonstrating an increase in activation in category-relevant visual areas (Folstein et al., 2013). Thus, the link between occipital activation and decision making we observe may be due to the engagement of neural representations tuned to diagnostic visual dimensions. It follows that this neural tuning may be most helpful and, therefore, recruited when generalizing to new stimuli.

Although the inclusion of the mid cingulate cortex (MCC) also improves DDM predictions, its potential role in this category learning paradigm is less clear. The relationship between MCC activation and drift rate was restricted to trained items, such that lower activation was associated with higher drift rates. While this relationship resembles that of the one observed in the LPFC, perhaps suggesting a role in encoding category rules during training trials, the lack of any association for testing trials is puzzling. In the absence of feedback, it may be that the MCC serves as a mere conduit between posterior and anterior cortical regions (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011).

A comprehensive account of category learning requires an understanding of the dynamics of attention, representation, and decision making both at the level of computational and neural processes. Here, we build on recent advances that link computational model predictions of category representations to neural coding (Mack, Love, & Preston, 2018; Zeithamova et al., 2019) to isolate the neural signals of category decision making. We also extend methods of linking brain and behaviour through the DDM (Frank et al., 2015; Mack & Preston, 2016; Roberts & Hutcherson, 2019; White, Mumford, & Poldrack, 2012) to demonstrate that trial-by-trial neural signals from occipital, mid cingulate, and lateral PFC track the accumulation of evidence in category decisions. Importantly, LPFC activation tracked participant-specific predictions of exemplar-based category evidence as formalized by EBRW (Nosofsky & Palmeri, 1997). More generally, this approach offers a novel method for quantitatively connecting behavioural data to neural processes with cognitive theory.

Acknowledgments

The authors thank Meg Schlichting for helpful discussions and Alison Preston and Bradley Love for providing access to the dataset. This work was supported by the Natural Sciences and Engineering Research Council (Discovery Grant RGPIN-2017-06753 to M.L.M) and the Canada Foundation for Innovation and Ontario Research Funds (JELF 36601 to M.L.M.).

References

Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences*, Vol. 22, pp. 170–188.

Bowman, C., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *The Journal of Neuroscience*, *28*, 2605–2614.

Braunlich, K., & Love, B. C. (2019). Occipitotemporal representations reflect individual differences in conceptual knowledge. *Journal of Experimental Psychology: General*, *148*, 1192–1203.

Braunlich, K., & Seger, C. (2016). Categorical evidence, confidence, and urgency during probabilistic categorization. *NeuroImage*, *125*, 941–952.

Davis, T., Goldwater, M., & Giron, J. (2017). From Concrete Examples to Abstract Relations: The Rostrolateral Prefrontal Cortex Integrates Novel Examples into Relational Categories. *Cerebral Cortex*, *27*, 2652–2670.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116.

Folstein JR, Palmeri TJ, & Gauthier I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*, 814–823.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V, Cavanagh, J. F., & Badre, D. (2015). FMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience*, *35*, 485–494.

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to

predictive coding in visual cortex. *Nature Neuroscience*, *19*, 665–667.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *The Journal of the Optical Society of America: A*, *20*, 1434–1448.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 13203–13208.

Mack, M. L., Love, B., & Preston, A. (2018). Building concepts one episode at a time: the hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38.

Mack, M. L., Preston, A., & Love, B. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*, 2023–2027.

Mack, M. L., & Preston, A. R. (2016). Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. *NeuroImage*, *127*, 144–157.

Medin, D., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Mendoza-Halliday, D., & Martinez-Trujillo, J. (2017). Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nature Communications*, *8*.

Mumford, J., Turner, B., Ashby, F., & Poldrack, R. (2012). Deconvoling BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*, 2636–2643.

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M., & Palmeri, T. J. (1997). An Exemplar-Based Random Walk Model of Speeded Classification. *Psychological Review*, *104*, 266–300.

O'Bryan, S. R., Walden, E., Serra, M. J., & Davis, T. (2018). Rule activation and ventromedial prefrontal engagement support accurate stopping in self-paced learning. *NeuroImage*, *172*, 415–426.

O'Bryan, S. R., Worthy, D. A., Livesey, E. J., & Davis, T. (2018). Model-based fMRI reveals dissimilarity processes underlying base rate neglect. *ELife*, *7*, 1–23.

Paniukov, D., & Davis, T. (2018). The evaluative role of rostrolateral prefrontal cortex in rule-based category learning. *NeuroImage*, *166*, 19–31.

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

Roberts, I. D., & Hutcherson, C. A. (2019, July 1). Affect and Decision Making: Insights and Predictions from Computational Models. *Trends in Cognitive Sciences*, Vol. 23, pp. 602–614. Elsevier Ltd.

Schaefer, A., Kong, R., Gordon, E., Laumann, T., Zuo, X., Holmes, A., … Yeo, B. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*, 3095–3114.

Seger, C., & Miller, E. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203–219.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–245.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *64*, 583–616.

Thompson-Schill, S. L., Swick, D., Farah, M. J., D'Esposito, M., Kan, I. P., & Knight, R. T. (1998). Verb generation in patients with focal frontal lesions: A neuropsychological test of neuroimaging findings. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 15855–15860.

Thompson, W. H., Wright, J., & Bissett, P. G. (2020). Open exploration. *ELife*, *9*. https://doi.org/10.7554/eLife.52157

White, C. N., Mumford, J. A., & Poldrack, R. A. (2012). Perceptual criteria in the human brain. *The Journal of Neuroscience*, *32*, 16716–16724.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*, 665–670.

Young, G., & Householder, A. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*, 19–22.

Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., van Kesteren, M. T. R., & Wutz, A. (2019). Brain Mechanisms of Concept Learning. *The Journal of Neuroscience*, *39*, 8259–8266.