A Varying Role for Abstraction in Models of Category Learning Constructed from Neural Representations in Early Visual Cortex

J. Brendan Ritchie and Hans Op de Beeck

Abstract

■ The human capacity for visual categorization is core to how we make sense of the visible world. Whereas a substantive body of research in cognitive neuroscience has localized this capacity to regions of human visual cortex, relatively few studies have investigated the role of abstraction in how representations for novel object categories are constructed from the neural representation of stimulus dimensions. Using human fMRI coupled with formal modeling of observer behavior, we assess a wide range of categorization models that vary in their level of

abstraction from collections of subprototypes to representations of individual exemplars. The category learning tasks range from simple linear and unidimensional category rules to complex crisscross rules that require a nonlinear combination of multiple dimensions. We show that models based on neural responses in primary visual cortex favor a variable, but often limited, extent of abstraction in the construction of representations for novel categories, which differ in degree across tasks and individuals. ■

INTRODUCTION

Categorizing what we see is crucial to how we make sense of the visual world. Investigating the neural representations that underlie this capacity has been a central focus of research in cognitive neuroscience on object categorization (for reviews, see Gauthier & Tarr, 2016; Grill-Spector & Weiner, 2014). However, one issue that has been underexplored is the role of abstraction in the neural representations that underlie learning of novel visual object categories.

The notion of abstraction is a key concept in models of category learning and is illustrated by the contrast between prototype and exemplar models. These models characterize the internal representation of stimulus dimensions in terms of a psychological space with the representation for stimuli occupying different points (Shepard, 1964; Attneave, 1950). Model representations for stimulus categories are then defined based on different distance functions applied between points in the space (Ashby & Maddox, 1993). Prototype models identify category representations with a representation of the central tendency of the distribution of individual stimulus representations (Smith & Minda, 1998; Reed, 1972; Posner & Keele, 1968). This summary representation abstracts away from the representation of individual stimuli, and when a novel stimulus is encountered, it is compared with the prototype for different categories. In contrast, exemplar models

involve no abstraction (Kruschke, 1992; Nosofsky, 1986, 1989); when categorizing a stimulus representation, it is compared with all known category members represented in the psychological space (Figure 1A).

Adjudicating between these models has been a source of considerable debate in cognitive science (e.g., Nosofsky & Zaki, 2002; Smith & Minda, 2000). However, despite overlap between research on visual object recognition and formal models of categorization (Palmeri & Gauthier, 2004; Op de Beeck, Wagemans, & Vogels, 2003), few neuroscientific studies have evaluated these models and, in some cases, focused exclusively on the predictions of one type of model (Davis & Poldrack, 2014; Op de Beeck, Wagemans, & Vogels, 2008). One important exception is Mack, Preston, and Love (2013) who compared the representations of both exemplar and prototype models with the neural patterns throughout the brain and found that the exemplar model provided better correspondence to neural responses in multiple visual brain regions.

The relatively modest amount of attention to the issue of abstraction is notable, because the frequent application of multivariate pattern analysis (MVPA) in neuroimaging experiments often rests on similar theoretical assumptions as prominent models of categorization. Indeed, when using MVPA, researchers typically assume that the brain represents stimuli via distributed patterns of neural activity, which when characterized geometrically, are distinguished by their distance in neural activation space (Haxby, Connolly, & Guntupalli, 2014;

Kriegeskorte & Kievit, 2013). Thus, in principle, one can apply the same distance functions that define different category representations to neural spaces reflecting the representation of stimulus dimensions (Op de Beeck et al., 2008).

Cognitive science has also begun to move away from the dichotomy between exemplar and prototype theories (Briscoe & Feldman, 2011; Vanpaemel & Storms, 2008; Love, Medin, & Gureckis, 2004). These two models can be seen as reflecting the extremes along a continuum of possible category representations, differing in the amount of abstraction, as defined by the different partitions of a stimulus set (Figure 1B). Recent neuroimaging studies have started to explore this larger range of possible category representations using an adaptive clustering model to characterize the neural signatures of rule-plus-exception category learning, which requires observers to form representations for new categories but then generalize to surprising test items. These studies have found that models that leverage degrees of abstraction, and are dynamically updated during learning, can be connected to representations in the medial temporal lobe (Davis, Love, & Preston, 2012) and the hippocampus (Mack, Love, & Preston, 2016; Davis et al., 2012).

Building on these theoretical and empirical developments, we sought to investigate the distance functions that the brain might apply to neural representations of stimulus dimensions, taking into account the continuum in levels of abstraction. Using a varying abstraction model (VAM) approach (Vanpaemel & Storms, 2008, 2010; Vanpaemel, Storms, & Ons, 2005), we applied multiple categorization models, which differ in their combination of exemplar and (sub)prototype representations, to a neural space for stimulus dimensions constructed from the human fMRI response in early visual cortex, which was then used to predict both group and individual choice behavior frequencies for four categorization tasks. Our results support a variable role for abstraction in the

representations of novel object categories derived from the neural representation of stimulus dimensions.

METHODS

Participants

The fMRI experiment included 11 adult volunteers (eight women; mean age = 25 years). The data of one participant were excluded from the analysis because of image distortions caused by a hair clip. Ten other adult volunteers participated in an offline similarity judgment experiment (seven women; mean age = 24 years). The categorization experiment included 80 volunteers (65 women; mean age = 20 years), with participants randomly assigned to one of four categorization task conditions (described below), resulting in 20 participants per task. All participants were predominantly right-handed, with normal or corrected vision, and provided informed consent to participate in the experiments. All experiments were approved by the ethics committee of KU Leuven.

Stimuli

Stimuli consisted of 16 annual square-wave gratings varying in orientation and spatial frequency (SF), centered on fixation and extending from 1.5° to 8° eccentricity in degrees of visual angle, with sharp interior and exterior edges. Gratings were presented at four orientations that differed in 30° steps (45°, 75°, 105°, and 135°) across four SFs, which varied in logarithmic steps (0.25, 0.5, 1, and 2 cycles per degree). In parametric space, this stimulus design produces a 4 × 4 grid pattern (Figure 2). For the experiments outside the scanner, a chin rest was used to ensure that viewing distance, and visual angle, was consistent across participants and experiments. Stimulus generation and control for all experiments in the study were via PCs running the Psychophysical Toolbox package

Figure 1. Possible category representations. (A) Exemplar and prototype category representations derived from stimulus space. (B) Five possible category representations, including three intermediary, as suggested by the VAM. Black dots indicate subprototypes in each category representation, whereas rings indicate stimulus representations fused into a subprototype.

Figure 2. Category structure for the four category learning tasks (dimensional, diagonal, interior–exterior, and crisscross). Boxes reflect an idealized stimulus space in which the distances between all stimulus points are equal along the dimensions of orientation (polar angle) and SF. Red and blue dots indicate stimuli that were assigned to Categories A and B, respectively. These stimuli were presented during the training phase of the category learning experiment. Rings indicate the unassigned stimuli added during the test phase. Numbers in the boxes indicate stimulus numbers (1–16).

(Brainard, 1997), along with custom code, in MATLAB (The Mathworks, Inc.).

The stimuli were selected because orientation is known to be readily decodable from early visual cortex (Haynes & Rees, 2005; Kamitani & Tong, 2005). These stimuli also allow for similar parameterization as so-called “Shepard circles,” which vary in orientation and size, resulting in a close link between our stimuli and a classic design frequently used in the categorization literature (e.g., Ashby & Maddox, 1990; Nosofsky, 1989; Shepard, 1964). Thirty-degree steps in orientation were selected because they were a decodable difference (Wardle, Ritchie, Seymour, & Carlson, 2017; Kamitani & Tong, 2005) and are the maximum difference that would allow for treating polar angle as a linear dimension. The logarithmic steps for SF were selected because of the nonlinearity of the cortical magnification factor as reflected in the BOLD response to stimuli of different SFs in V1 (Harvey & Dumoulin, 2011; Henriksson, Nurminen, Hyvarinen, & Vanni, 2008) and so were expected to produce more linear steps in neural space. Although studied less frequently, we expected SF to be readily decodable from early visual cortex, and it was chosen instead of size as the second dimension because piloting suggested it produced variations in neural response that were more equitable to those for differences of orientation.

Similarity Experiment

Participants were informed they would be required to judge how similar stimuli were with respect to the orientation and “thickness” of grating bars (i.e., SF). Participants responded using a keyboard number pad based on a 6-point scale (1 = *highly similar*, 6 = *highly dissimilar*). At the beginning of the experiment, a preview was provided of each stimulus dimension so that participants could briefly cycle through the different levels of one or the other stimulus dimension (the orthogonal dimension

value during the previews was not used in the main experiment; e.g., 0° orientation when previewing SF). After the preview, all the stimuli were cycled through once (2500-msec stimulus duration) in random order, and then the task began. On each trial, a prestimulus fixation period (500 msec) was followed by a first stimuli (1000-msec duration), an ISI in which only the fixation point was displayed (1000 msec), and then the second stimulus (1000 msec), after which text appeared reminding participants of the 6-point scale. The next trial did not begin until a response was recorded. All sequential pairwise combinations of stimuli were presented during the experiment, with grating phase varying randomly between the two phases presented during the fMRI scanning procedure (described below).

Categorization Experiment

Each categorization task had the same design. During two training phases, participants learned to guess the category membership of a subset of category assigned stimuli and generalize this information during a test phase in which the remaining stimuli were presented. This three-stage structure was based on classic categorization studies (e.g., Nosofsky, 1987). On each trial in the first training phase, a stimulus was presented after a prestimulus fixation period (500 msec) until participants made a guess, and then feedback was provided in the form of text indicating whether the response was correct and the category label of the stimulus. The stimulus remained on the screen during the feedback, and feedback information remained on the screen until participants pressed the space key for the next trial to begin. For the second training phase, stimuli were presented for 200 msec after the prestimulus period, followed by a 1750-msec response period, after which feedback was provided at fixation, where a bull’s eye would flash green if the response was correct and red if incorrect (250 msec). Each of the assigned (8/16)

stimuli were presented 16 times in each of the training phases. During the test phase, all 16 stimuli were presented eight times in the same format as the second training phase, and feedback on performance was still provided for the assigned stimuli. As with the similarity task, grating phase varied randomly between the two phases from the fMRI study.

At the beginning of the experiment, participants were instructed that they would be performing a task that required them to learn to discriminate between simple objects from two categories. After the three stages of the task were described to the participants, they performed a brief toy experiment (using numeral stimuli) to familiarize them with the overall task design before beginning the main experiment.

Four stimuli each were assigned to one of two categories for each task. The category structures of the four categorization tasks were as follows (Figure 2). First, in the “dimensional” task, stimuli were grouped based on their values along the dimension of SF. Second, in the “diagonal” task, stimuli were grouped so that they varied along both the orientation and SF dimensions. Third, in the “interior–exterior” task, the stimuli were grouped into either predominantly middle values along the two stimulus dimensions or more extreme values. Fourth, in the “crisscross” task, stimulus categories were bimodal along both stimulus dimensions. Importantly, unlike the first two tasks, the latter two involved categories that are not linearly separable in parametric space.

The category structures for the four tasks were all based on experiments from the categorization literature that have been analyzed using VAM (Vanpaemel & Storms, 2010). In particular, the crisscross task was included because simulations have shown this structure to be especially good for differentiating between different category representations with VAM (Vanpaemel & Storms, 2008). In contrast, the dimensional task was included as a baseline condition, where we expected all models to fit well to the categorization choice frequency data.

For each task, the dependent variable of interest for subsequent modeling with VAM was the total frequency of Category A choice responses across participants for each stimulus, which is standard in the categorization literature, including the 30 data sets reanalyzed by Vanpaemel and Storms (2010). For this analysis, the data of individual participants were only included if they performed >61% accuracy at the test phase of the experiment and were otherwise excluded from further analysis. This threshold is similar to those employed for excluding participants on difficult tasks in the categorization literature (Nosofsky, 1987) and conforms to the lower bound of frequency of correct responses that is significantly different from chance based on a one-tailed binomial test ($\alpha = .05$). If participants performed below this threshold, their data were not included in the modeling analysis.

To test the reliability of the data for the participants who performed above the accuracy threshold, we com-

puted the Pearson r correlation between the individual Category A response frequencies and the average frequencies from the remaining participants. The significance of the average correlation across folds of this leave-one-out procedure was then tested for significance using a permutation test (1,000 iterations). For the permutation test, the p value describes the proportion of iterations with an effect size greater than the observed correlations.

Scanning Procedure

Scanning consisted of a single session of 12 experimental runs lasting 3 min 8 sec, followed by an anatomical scan. Runs were designed to be brief because this has been shown to improve neural pattern discriminability with human fMRI (Coutanche & Thompson-Schill, 2012). Each run consisted of two random sequences of 16 stimulus and five fixation trials for a total of 42 trials per run. Each stimulus trial began with the stimulus being presented for 2000 sec, phase reversing at 4 Hz, followed by 2000 msec of fixation. This resulted in 24 presentations of each stimulus throughout the experiment. During scanning, participants were required to judge how similar the current stimulus was to the previous one (a 1-back similarity task) based on a 4-point similarity scale. Responses were made using the middle and index fingers from both hands (left middle = *very similar*, right middle = *not very similar*).

Imaging Parameters

Data acquisition was carried out using a 3-T Philips scanner with a 32-channel coil at the Department of Radiology of the KU Leuven university hospitals. Functional MRI volumes were acquired using a 2-D T2*-weighted EPI sequence: repetition time = 2 sec, echo time = 30 msec, flip angle = 90°, field of view = 216, voxel size = $3 \times 3 \times 3$ mm, and matrix size = 72×72 . Each volume consisted of 37 axial slices aligned parallel to the calcarine sulcus of each participant, with no gap. The T1-weighted anatomical volumes were acquired with a magnetization prepared rapid gradient echo sequence, with a $1 \times 1 \times 1$ mm resolution.

fMRI Preprocessing and Analysis

Preprocessing and analysis of the MRI data were carried out using SPM12 (v.6906). For each participant, fMRI data were slice-time corrected, motion corrected, coregistered to the individual anatomical scan, and smoothed using a Gaussian kernel of 6-mm FWHM (Op de Beeck, 2010). No normalization was performed, and all analyses were carried out within the native brain space of the participant. After preprocessing, the signal for each stimulus, at each voxel, was modeled using a general linear model. The predictors for the model consisted of the 16 stimulus

conditions and six motion correction parameters (translation and rotation along the x , y , and z axes). The time course for each predictor was characterized as a boxcar function convolved with the canonical hemodynamic response function. This produced one parameter estimate for each voxel, for each predictor, for each run.

ROI Definition

The ROI for the study consisted of portions of the bilateral primary visual cortex (V1), because it is known to code for both orientation and SF. Furthermore, Mack et al. (2013) found that categorization model representations have a close correspondence to the neural responses from early visual cortex for both exemplar and prototype models. The ROIs were defined anatomically using the following procedure. First, cortical surface reconstruction was performed on the anatomical MRI scans of each participant using FreeSurfer 5.3 (Fischl, 2012). Second, the templates for eccentricity and area boundaries of Benson, Butt, Brainard, and Aguirre (2014) for V1 were separately registered to the inflated cortical surface of the hemispheres of each participant. This method has been found to be at least as reliable for demarcating the retinotopic structure of these regions as 10 min of scanning for standard retinotopy (Benson et al., 2012). Using this template, the boundaries of the region were defined, and then the area ROI was convolved with the eccentricity template, cropped to 1.5° – 9° . This resulted in bilateral ROIs for V1 (mean voxel number = 404, $SD = 39$) corresponding approximately to the portions of these regions that map to the retinotopic position of the stimuli.

Decoding Analysis

Decoding analysis was carried out using linear discriminant analysis classifiers as implemented in the CoSMoMVA toolbox (Oosterhof, Connolly, & Haxby, 2016), which were trained and tested on the general linear model parameter estimates (beta weights) of the ROI voxels for each of the 16 stimulus conditions.

We sought to investigate the pairwise discriminability of the stimuli based on the difference in levels along each of the stimulus dimensions using leave-two-run-out cross-validation, for all possible unique run pairs (66 in total). For the pairwise classification, we also sought to determine whether information about orientation and SF generalized across levels of the orthogonal dimension. We therefore combined leave-two-run out cross-validation with leave-one-level-out cross-validation—what we will call “run-plus-level” cross-validation. For this form of cross-validation, a target pairwise comparison was chosen, and data for these two conditions were selected from the two test runs. For the remaining 10 training runs, the data were selected for the three stimulus pairs that were equivalent along the target dimension but had values at one of the other three levels of the orthogonal dimension. This

allowed us to test the independent generalizability of orientation and SF information in the neural responses from the ROI.

The statistical significance of between-participant classification performance effects was tested with the non-parametric Wilcoxon signed-rank test. To assess the effect of differences in stimulus levels on pairwise classification performance, we performed repeated-measures ANOVA to reveal any effect of difference in level and fit linear mixed effects models to determine whether there was a linear increase of classification accuracy with an increase in stimulus difference along the two dimensions.

Representational Similarity Analysis

To construct a 2-D neural space for the stimulus dimensions, we used representational similarity analysis to make neural dissimilarity matrices (DMs) from the ROI neural response patterns (Kriegeskorte, 2008). Our measure of neural dissimilarity was the cross-validated Mahalanobis distance (Walther et al., 2016):

$$d_{ij}^2 = (v_i - v_j)_{TR} \Sigma_{TR}^{-1} (v_i - v_j)_{TS}^T$$

where v_i and v_j are row vectors for the neural pattern for the i th and j th stimulus conditions, TR and TS denote the training and test partitions of the data, and Σ_{TR} is the between-condition covariance matrix for the training data. This measure of dissimilarity is equivalent to the distance along the linear discriminant analysis discriminant because the Euclidean distance is normalized by the training sample covariance. Because the number of voxels (features) in an ROI may exceed the number of beta weight estimates (samples), Σ_{TR} may be rank deficient and thus noninvertible (Walther et al., 2016). To address this issue, we regularized the training sample covariance matrices using the oracle approximating shrinkage estimator (Chen, Wiesel, Eldar, & Hero, 2010).

Neural DMs were constructed using one of two procedures. First, for standard construction of a DM, each cell of the matrix was based on the pairwise distance averaged across cross-validation folds, again using a leave-two-run-out design. Second, we also sought to construct neural DMs in a manner that took account of the independence of the two stimulus dimensions. For this, we again performed run-plus-level cross-validation. This approach only allows for an estimate of pairwise relationships of neural patterns for stimuli that differ in value along one dimension. To estimate all distances in the matrix, we extrapolated the missing values in the matrix by computing the hypotenuse for the differences between two stimulus conditions along the two stimulus dimensions, which is analogous to the procedure for constructing a DM from stimulus parameter values.

To test for the reliability of the DMs for each ROI, we performed both within- and between-participant estimates of the reliability of the data (Ritchie, Bracci, &

Op de Beeck, 2017; Walther et al., 2016). For within participant, the data were portioned into odd and even runs, and the above two forms of cross-validated distances were computed, using a leave-one-run-out design. For the between-participant comparisons, a single neural DM was calculated for each participant using leave-two-run-out cross-validation, and the correlation was then computed between group-averaged DMs based on all permutations of an even split between participants. After testing for reliability, a group-averaged neural DM for V1 was constructed, which was correlated with the behavioral and parametric DMs as well as the DMs for other ROIs.

When testing for reliability within participants, a Pearson's r coefficient with a fixed threshold was used, denoted r_{fix} , which is sensitive to any difference in the average distance between two DMs and so reflects whether ratios of distance are consistent between matrices (Walther et al., 2016). Where multiple correlations were computed, significance was assessed using a permutation test (1,000 iterations). As before, permutation test p values reflect the number of iterations with an effect size greater than the observed effect.

DMs were also constructed from participants' similarity ratings and the stimulus parameterization, which were compared with the neural DMs. A behavioral DM was constructed for each participant based on his or her similarity ratings, which were rescaled to range between 0 and 1. For the offline task, the average response for each pairwise comparison (regardless of presentation order) was entered into the appropriate matrix cell. For the data collected in the scanner, a separate matrix was made for each run, with missing values replaced by the within-run average dissimilarity. The matrices were then averaged across runs for each fMRI participant. The relationship between the group-averaged DMs for the offline and scanner participants was assessed based on the Pearson's r correlation between the bottom half of the matrices, before both group matrices were combined into a single behavioral DM. For the parametric DM, we inverted the value of cycles per degree for SF, so that differences might more closely reflect the subjective differences of the stimuli, because perceptually a difference in 0.5 and 0.25 cycles per degree appears much greater than the difference between 1 and 2 cycles per degree. Thus, plausibly, the subjective similarity might more closely reflect the degrees per cycle (i.e., the "thickness" of the grating bars), which we will call SF^{-1} . The relationship between the group-averaged neural, behavioral, and parametric DMs was assessed using Spearman's ρ rank-order correlation.

Multidimensional Scaling

In our study, we aimed for an experimental design similar to that of psychological research modeling categorization choice behavior, but with a neural space acquired from

the BOLD response of participants (distinct from those performing the categorization tasks) taking the place of a psychological space constructed from similarity judgments (Op de Beeck et al., 2008). To generate a 2-D neural space from the group-averaged neural DMs, we performed nonmetric multidimensional scaling (MDS) with Stress 1 as the loss function. Interpretation of stress followed the recommendations of Kruskal (1964), with a level of 0.10 considered a threshold for a fair level of goodness-of-fit. In particular, our focus was on the stress for 2-D neural spaces, although MDS was also carried out for visualization of the group-averaged behavioral and parametric DMs. A neural space with low stress served as input to VAM for modeling observer choice frequencies.

VAM

VAM is motivated by the insight that prototype and exemplar models reflect the two extremes in the level of abstraction of category representations. Between these extremes, intermediate cases can be defined by different partitions of a stimulus set, which generates different sets of subprototypes of the category (Vanpaemel & Storms, 2008; Vanpaemel et al., 2005; Figure 1B). From this collection of representations, one can specify different models using the architecture of the generalized context model, the most well-known exemplar model (Nosofsky, 1986). These models differ solely in their category representations (i.e., degree of abstraction) and otherwise have identical parameterization.

Following generalized context model, the distance between a stimulus and a subprototype in a 2-D representational space is characterized by VAM as a weighted city-block metric:

$$d_{ia} = w|x_{i1} - x_{a1}| + (w - 1)|x_{i2} - x_{a2}|$$

where x_{im} is the value of the i th stimulus along dimension m , x_{am} is the value of the subprototype a of Category A along m , and w ($0 \leq w \leq 1$) is the attentional weight parameter for the dimensions. In turn, the similarity between the stimulus and the subprototype is an exponential decay function of their distance:

$$s_{ia} = e^{-cd_{ia}}$$

where c is a sensitivity parameter. The probabilities of making an A response to the stimuli are then modeled by applying the Luce choice rule (Luce, 1959) to the overall similarities between the stimulus and the two Categories A and B. In the present version of VAM, the overall similarities are the result of summing and then scaling, by the number of subprototypes, the pairwise similarity between the stimulus and category subprototypes (i.e., the mean similarity):

$$R_{iA} = \frac{\sum_{a \in A} s_{ia} / N_A}{\left[\sum_{a \in A} s_{ia} / N_A \right] + \left[\sum_{b \in B} s_{ib} / N_B \right]}$$

where A and B are the sets of category subprototypes and N_A and N_B are the number of elements in A and B , respectively. In all previous versions of VAM, the category level similarity is summed, but not scaled. However, this potentially penalizes models that have a substantial difference in level of abstraction between categories (e.g., an exemplar representation for one category and a prototype for the other). For this reason, we have chosen to scale the summed similarities by the number of subprototypes.

Under this version of VAM, there are only two free parameters: the sensitivity parameter c and the attentional weight parameter w . The response bias parameter, b , and the response scaling parameter, γ , were not free parameters in this implementation of VAM and were fixed to 0.5 and 1, respectively. For simplicity, they are therefore excluded from the above statement of the choice rule.

Categorization Model Selection

Under VAM, the number of possible categorical representations, and models, is determined by the number of partitions of a set—also known as the Bell number. For our categorization tasks, the number of possible models is determined by multiplying the Bell numbers of the sets of stimuli assigned to Categories A and B: $\text{Bell}(4) \times \text{Bell}(4) = 225$. However, not all models are necessarily psychologically plausible. To reduce the number of models, we performed hierarchical clustering on each set of the assigned stimuli based on the V1 neural DM. Only models with subprototypes corresponding to a level in the clustering hierarchy were selected for our analysis (the exemplar and prototype models always correspond to the bottom and top of the hierarchy), which reduced the number of models to the number of stimuli in each category (Verbeemen, Vanpaemel, Pattyn, Storms, & Verguts, 2007). Thus, for each categorization task, we only included $4 \times 4 = 16$ models in the analysis, with one for every unique combination of the four possible representations generated for each of the two categories based on the levels of the hierarchical clustering.

Parameter Estimation and Model Fitting

Modeling of group- and individual-level categorization task performance was carried out using Markov chain Monte Carlo simulations implemented in MATLAB and JAGS (Plummer, 2003), with code adapted from previous studies (Bartlema, Lee, Wetzels, & Vanpaemel, 2014; Lee & Wagenmakers, 2014; Vanpaemel, 2009). The observed response frequencies were modeled as

$$y_i \sim \text{Binomial}(R_{iA}, t)$$

where R_{iA} is the observed frequency of responding that the i th stimulus is a member of Category A and t is the

number of stimulus presentations during the test phase of the categorization tasks (summing across participants for the group level fits). The priors for the two free parameters were as follows:

$$w \sim \text{Uniform}(0, 1)$$

$$c^2 \sim \text{Gamma}(e, e)$$

where $e = 0.001$. The justification of the prior for c follows the reasoning of Vanpaemel (2009). Because $1/c$ scales the distances between stimulus points (Op de Beeck et al., 2008), the parameter serves as an inverse scale, which suggests that c^2 operates as a precision. The gamma distribution is the standard noninformative prior for precision and has the desirable property that it makes any claim about c near invariant to the scaling of distance.

Three Markov chains were run for 10^5 iterations for the group fits and 10^4 iterations for the individual fits, each with 10^3 -step burn-in. The rank order of model fit was assessed by calculating $-2\ln(L)$, which is the minimum deviance (or $-2 * \ln$ of the maximum likelihood estimate) based on the Markov chain Monte Carlo simulations. Pairwise differences in minimum deviance between models were compared via both Bayesian methods and inferential statistics. First, the difference between values of $-2\ln(L)$ roughly approximates the Bayes factor when making pairwise comparisons between models (Verbeemen et al., 2007) and was interpreted based on the prescriptions of Kass and Raftery (1995) for the Bayes factor: 0–2 = negligible evidence, 2–6 = positive evidence, 6–10 = strong evidence, and >10 = very strong evidence. Second, the difference in deviance was tested for significance using a chi-square test (cf. Mack et al., 2013), because deviance is known to have a chi-square distribution. Because all models were identical in parameterization, it was not necessary to compare model fits based on measures (such as the different informational criteria) that are sensitive to differences in degrees of freedom. For the group fits, to assess how well the predicted probabilities explained the observed variance in response frequencies, the frequencies were rescaled and R^2 and Pearson's r were also calculated between the observed and predicted response proportions for models that were directly compared.

RESULTS

Behavioral Performance on Categorization Tasks

Eighty participants participated in a categorization experiment and performed one of four categorization tasks. At the test phase of the experiment, most participants performed well above the 61% accuracy threshold (dimensional: 20/20; diagonal: 19/20; interior–exterior: 13/20; crisscross: 18/20; Figure 3). The data of participants who performed below threshold were excluded from all

Figure 3. Observer choice accuracies for the categorization tasks. Boxplots show the median accuracy for the four tasks (DM = dimensional; DI = diagonal; IE = interior-exterior; CC = crisscross). Black rings are for participants who performed above the inclusion threshold (61% accuracy; red hash line), and gray rings indicate those participants who performed below threshold. Note that, because in some cases participants had identical accuracies, a single ring may reflect the responses of multiple participants (e.g., for the dimensional task, only seven rings appeared, although there were 20 participants).

further analyses. Mean choice accuracy for the participants who performed above threshold was highest for the dimensional task (mean = 96% accuracy), with similar mean accuracies for the diagonal and crisscross tasks (mean = 83% and 85%, respectively). Performance was lowest for the interior-exterior task (mean = 76%), for which the categories appear to have been the most difficult for participants to learn.

To verify that the general pattern of response frequencies was relatively reliable across participants, we performed a leave-one-out procedure correlating individual choice frequencies with the remaining participant group average. These correlations were extremely high for the dimensional (mean $r = .99$) and crisscross (mean $r = .90$) tasks and to a lesser extent for the diagonal (mean $r = .74$) and interior-exterior (mean $r = .62$) tasks. All of these correlations were highly significant when assessed using a permutation test ($p = 0$ for all tasks). Figure 4A depicts, for each task, the Category A choice frequencies for the individual stimuli and plots both the group means (rings) and individual data points (gray lines). For the dimensional task, choice probabilities were highly consistent for both the assigned and unassigned stimuli. For the other tasks, performance was more variable across participants. This was especially true for the participants who performed the interior-exterior task. When the group-averaged response frequencies are plotted as color blocks in regions of stimulus space, it can be seen that they generally reflect natural patterns of generalization to the unassigned stimuli (Figure 4B).

Decoding Stimulus Orientation and SF from Early Visual Cortex

Our stimuli varied along two dimensions, orientation and SF; therefore, we sought to decode information about these dimensions from neural patterns in V1 and evaluate classification accuracy as a function of the pairwise differences between stimuli along the two dimensions. For the neural responses from V1 to serve as a useful measure for a 2-D activation space, they must have two key

Figure 4. Choice frequencies for the categorization tasks. (A) Rings indicate the group-averaged proportion of Category A responses for the four categorization tasks, for each of the 16 stimuli. Red and blue rings indicate whether a stimulus was assigned to Category A or B (respectively), and black rings indicate unassigned stimuli. Gray lines indicate the response frequencies across stimuli for an individual participant. (B) Group-averaged Category A response proportions plotted as regions of an idealized stimulus space. Red indicates proportions > 0.5 and a greater number of Category A responses, whereas blue indicates proportions < 0.5 and a greater number of Category B responses. Stimulus numbers are located in the center of each of the regions.

Figure 5. Pairwise decoding of stimulus orientation and SF from V1. (A) Matrices show classifier accuracy plotted to reflect increased differences in either orientation or SF (note the difference in scale) based on standard leave-run-out cross-validation. Boxplots show median classifier performance across differences in orientation and SF between stimuli, collapsing across cells in the matrices. Gray lines and rings indicate data from an individual participant ($N = 10$). (B) Results using run-plus-level cross-validation, as described in the text (note again the difference in scale for the matrices).

properties. First, the responses for stimuli that minimally differ along the two stimulus dimensions must nonetheless be discriminable; otherwise, the characterization of the 2-D space will be too coarse. Second, pairwise decoding should show a linear increase as parametric difference increases. Thus, we sought to verify that our data had these two properties.

We performed pairwise classification of the neural patterns of stimuli focusing on pairs within the same level of the orthogonal dimensions and then averaged classifier performance across levels. This provided measures of classification performance for differences of 30°, 60°, and 90° orientation and one, two, and three levels of SF (Figure 5A). Classification accuracy was significantly different from chance for all differences levels for both orientation (for all differences: $W(9) = 55$, $z = 2.803$, $p = .005$) and SF (Level 1: $W(9) = 55$, $z = 2.807$, $p = .005$; Level 2: $W(9) = 55$, $z = 2.803$, $p = .005$; Level 3: $W(9) = 55$, $z = 2.812$, $p = .005$). Repeated-measures ANOVA revealed highly significant main effects of difference level on classifier performance for both orientation, $F(2, 18) = 15.208$, $p = 1.35e-4$, and SF, $F(2, 18) = 100.14$, $p = 8.69e-9$. We next fit linear mixed effects models to the classification accuracies for differences in orientation and SF. These models were also highly significant (orientation: $t(28) = 4.698$, $p = 6.32e-05$, $R^2 = .82$; SF: $t(28) = 11.605$, $p = 3.27e-12$, $R^2 = .87$) and showed that there was a linear increase in classifier accuracy with an increase in difference along stimulus dimensions (Figure 5A; Wardle et al., 2017; Kamitani & Tong, 2005). We note that, with respect to pairwise relationships, only the comparison between differences of 60° and 90° was not significant, $W(9) = 31$, $z = 0.358$, $p = .721$.

We also sought to determine whether information about orientation and SF could be decoded from V1 when, in the analysis, we assumed the dimensions were coded independently (Figure 5B). To this end, we again performed the above analyses, now using run-plus-level cross-validation. Although overall classifier performance was generally slightly lower, decoding was again significantly different from chance for all difference levels for both orientation (for all differences: $W(9) = 55$, $z = 2.803$, $p = .005$) and SF (for all differences: $W(9) = 55$, $z = 2.803$, $p = .005$). We again performed repeated-measures ANOVA, which revealed main effects of difference level on classifier performance for both orientation, $F(2, 18) = 19.814$, $p = 2.83e-05$, and SF, $F(2, 18) = 147.69$, $p = 6.81e-012$. We then again fit linear mixed effects models to the classification accuracies based on the differences along the two dimensions. As before, these models were highly significant (orientation: $t(28) = 5.492$, $p = 7.23e-06$, $R^2 = .88$; SF: $t(28) = 13.285$, $p = 1.30e-13$, $R^2 = .94$), showing a linear increase in classification accuracy with increased difference in orientation and SF. This shows that information about orientation and SF in V1 transcends levels of the orthogonal stimulus dimension, which supports the contention that these dimensions are independently coded for in V1. Again, the comparison between 60° and 90° was not significant, $W(9) = 46$, $z = 1.886$, $p = .059$.

A Low Dimensional Neural Space for Orientation and SF

We sought to construct a group-averaged neural DM from neural responses in V1 from which to generate a low-dimensional neural space as input to VAM to model

observer choice frequencies. To this end, we first investigated the reliability of the dissimilarity relations between the neural patterns for different stimulus conditions. We first tested these relationships within participants, by constructing cross-validated distance matrices for split-half data and looking at the within-participant correlations (Walther et al., 2016). We found that the within-participant split-half data were well correlated ($r_{\text{fixed}} = .29$, $SD = .10$), $W(9) = 55$, $z = 2.803$, $p = .005$, suggesting that the data of individual participants were reliably similar over the course of the experiment. Next, we evaluated the reliability at the group level by constructing individual participant DMs for V1 and then computed the average of all permutations of the split-half correlations across participants. Again, the neural RDMs were highly correlated (mean Pearson's $r = .89$) and significant based on a permutation test ($p = 0$).

We compared the group-averaged DM for V1 to the behavioral DM from individual similarity judgments as well as a DM based on the stimulus parameterization (Figure 6). The group-averaged behavioral DMs for the offline and scanner participants were highly correlated ($r = .91$, $p < 8.54e-48$) and therefore combined for this analysis. Both matrices were highly correlated with the neural DM (parametric DM: $\rho = .76$, $p = 3.94e-24$; behavioral DM: $\rho = .74$, $p = 5.53e-22$). Using these DMs, we then performed 2-D MDS. The resulting parametric space was as anticipated, with nonlinear separation along

the SF^{-1} dimension (due to the logarithmic differences in SF between stimuli), although with poor stress (Stress 1 = 0.21). For the 2-D behavioral space, notable was a “left–right” division, although stress was again poor (Stress 1 = 0.20). For the V1 pairwise neural DM, we observed a space in which distances were much greater along the SF dimension, which is consistent with the higher classifier accuracy for differences in SF than orientation. The stress for this space was quite low (Stress 1 = 0.06); however, it included distortions relative to the predicted relative position of points for individual stimuli. Specifically, Stimulus pairs 9–10 and 13–14 were reversed in spatial position relative to their parameterization for orientation. We therefore sought to construct an extrapolated neural DM for V1, using run-plus-level cross-validation, which might lack these local imperfections (Figure 6A).

As before, we tested the within- and between-participant reliability of the extrapolated neural DMs. The within-participant split-half matrices were incredibly well correlated with each other (mean $r_{\text{fixed}} = .9$, $SD = .06$), $W(9) = 55$, $z = 2.803$, $p = .005$, which likely reflects the greater amount of averaging (and greater interdependence of values) of data for the within-participant DMs. The mean group split-half correlation was also very high (mean $r = .93$), again likely reflecting the increased amount of averaging in the construction of the extrapolated DMs, and was also significant as determined by a permutation test ($p = 0$). In addition, as with the pairwise matrix, the extrapolated

Figure 6. Results of representational similarity analysis and MDS. (A) DMs based on stimulus parameterization, behavioral similarity judgments, and pairwise cross-validated Mahalanobis distances for V1. Note that, for the parametric DM, differences reflect SF^{-1} (see main text). Two-dimensional spaces constructed from the matrices using MDS (Stress 1) are also depicted. Black circles indicate stimulus points, accompanied by stimulus numbers. (B) Correlation matrix comparing the parametric (param), behavioral (behav), pairwise distances from V1 (V1-p), and extrapolated DM from the region (V1-e). (C.) Extrapolated DM based on the run-plus-level cross-validated Mahalanobis distances (see main text). Also depicted is a 2-D neural space generated by applying MDS to the extrapolated DM.

neural DM was well correlated with both the parametric ($\rho = .73, p = 3.78\text{e-}21$) and behavioral ($\rho = .74, p = 8.16\text{e-}22$) DMs. The two neural DMs were, as expected, very highly correlated with each other ($\rho = .95, p = 2.58\text{e-}61$). However, crucially, the resulting 2-D space constructed from the extrapolated neural DM showed no local disorder of position (Figure 6C) and had even lower stress (Stress 1 = 0.03). Thus, the extrapolated DM can be thought of as a more regularized DM because it relies on more robust estimates of key dissimilarities along each independent dimension and extrapolates to remaining dissimilarity relations to generate the full matrix. This 2-D space served as input to our modeling of observer choice frequencies.

Modeling Group Categorization Behavior Using a Neural Space of Stimulus Dimensions

The 2-D V1 neural space was used to calculate the similarity metric for VAM and estimate the predicted response probabilities for each stimulus of the 16 different models for each of the four different categorization tasks. The fits for these models, as $-2\ln(L)$, are depicted in Figure 7A. The vertical axes reflect the minimal deviance for each model, with lower values indicating better model fit. The numbers along the horizontal axes reflect the number of subprototypes in each category model. Thus, the 4-4 bar depicts the fit for the exemplar model, whereas the 1-1 bar depicts the fit for the full prototype model. All other bars reflect models intermediary in their level of abstraction. We will summarize the modeling results for the four tasks in turn (Table 1).

First, for the dimensional task, several models fit extremely well. All of these models were close to perfect at accounting for the variance in the data and give complete attentional weight to the SF dimension, as would be expected given the category structure of the task. Four of these models, including the exemplar model, had a virtually identical fits, whereas the prototype model fits slightly worse. In particular, the fits of the best intermediate models (2-4, 4-2, and 2-2) were identical to that of the exemplar model, up to a rounding error. The difference in fit between all of these models and the fit of the prototype model ($86.55 - 66.09 = 20.46$) suggests very strong evidence in favor of these models, which was also supported by a test of significance, $\chi^2(1) = 20.46, p = 6.097\text{e-}06$.

Second, for the diagonal task, the best fitting model was an intermediate model (Model 4-3), which was slightly better than the exemplar model. The difference ($268.16 - 265.19 = 2.96$) provided some positive evidence for the intermediate model but was not significant, $\chi^2(1) = 2.96, p = .083$. Both the best intermediate and exemplar models offered substantially better fits than the prototype model, which was second worst of the 16 models tested (Figure 7A). Although, in principle, both dimensions

require attention to learn the task, the values of w are <0.5 . This may be because participants attended more to the orientation dimension or a compensation on the part of the models for the fact that the orientation distances were generally smaller than those along the SF dimension.

Third, for the interior-interior task, the difference in fit ($226.02 - 209.14 = 16.88$) provided very strong evidence in favor of the exemplar model over the best intermediate model (3-4). This difference was significant, $\chi^2(1) = 16.88, p = 3.9819\text{e-}05$. Again, both models provided substantially better fits than the prototype model (Figure 7A). The exemplar model also clearly accounted for more of the variance in the response frequencies (Table 1). The superior good fit of the exemplar model may reflect the fact that participants found this task to be comparatively difficult and so perhaps relied more on a learning strategy that involved memorizing the individual stimulus items. Again, much greater weight was shown for the orientation than SF dimension (Table 1).

Fourth, for the crisscross task, the difference in fit ($473.29 - 406.5 = 66.79$) provided very strong evidence in favor of the best intermediary model (3-4) over the exemplar model, $\chi^2(1) = 16.88, p = 3.0203\text{e-}16$. The best intermediate model also accounted for more of the variance in the response frequencies (Table 1). In contrast, the prototype model provided an exceptionally poor fit to the data, which was expected given that two category prototypes would occupy virtually the same position in the V1 neural space. In accordance with this, the prototype model accounted for virtually none of the variance in the data. In contrast, the predicted response proportions of the best intermediary model accounted for a good deal of the variance in the observed proportions.

Figure 7B shows the best fitting model category representations, imposed on the 2-D neural space from V1. For the dimensional task, because there was a four-way tie, the most abstract model is depicted. To further illustrate the quality of the best fitting model, Figure 7C shows the predicted responses plotted against the observed response proportions and the best fitting model. As can be seen for all tasks, the predicted response proportions were closest for the assigned as opposed to the unassigned stimuli. This is also clear when differences in response proportions are plotted based on cells in the stimulus space (compare Figure 7D with Figure 4B). Notably, for the diagonal and crisscross tasks, the response proportions for some unassigned stimuli that were consistently categorized by observers were poorly captured by even the best models. For the diagonal task, this included responses for Stimuli 6 and 11, whereas for the crisscross task, this included Stimuli 3, 5, 12, and 14. These poor fits may be related to the fact that several assigned stimuli from both categories were adjacent to these stimuli in neural space, which may have had less of an impact on how observers generalized the stimuli. Taken as a whole, these modeling results provide

Figure 7. Results of modeling group-level categorization task choice frequencies using a neural varying abstraction model. (A) Three-dimensional bar plots of model fits for each of the 16 models for each of the four categorization tasks. Vertical axis reflects model fits in $-2\ln(2)$ (minimal deviance), with higher bars indicating worse fits. Horizontal axes reflect the number of subprototypes in each model (exemplar = 4-4, prototype = 1-1). (B) Best fitting model category representations for the data from each task, applied to the 2-D neural space from V1. Note that, for the dimensional task, there was a four-way tie, and the most abstract of these models is depicted. Solid red and blue circles indicate the subprototypes of models, and colored rings indicate the fused stimulus representations, whereas black rings indicate the unassigned stimuli. Number pairs in the top right corners indicate the number of subprototypes in the depicted representation coded by category (red = A, blue = B). (C) Predicted choice response properties of the best fitting model for each task (gray line) are plotted against the observed group-averaged response proportions. Other figure details are the same as in Figure 4A. (D) Predicted Category A response proportions of the best fitting models are plotted as regions of an idealized stimulus space. Other figure details are the same as in Figure 4B.

evidence of a modest degree of abstraction, depending on the categorization task learned by observers.

Modeling Individual Categorization Behavior Using a Neural Space of Stimulus Dimensions

The results of the group level fits suggest a modest role for abstraction. However, one possibility is that there are

between-participant differences in category learning, and levels of abstraction, which are being obscured by pooling the choice frequencies of individual observers. To investigate this, we fit the different models to the individual participant data for the four tasks. We then compared the best fitting model with some level of abstraction (including the prototype model) with the exemplar model, calculating the difference in $-2\ln(L)$ between the two

Table 1.

Task	Model	$-2\ln(L)$	R^2	r	c	w
Dimensional	Exemplar	66.09	.99	.99	0.17	1
	Best intermediate	66.09	.99	.99	0.17	1
	Prototype	86.55	.99	.99	0.16	1
Diagonal	Exemplar	268.15	.83	.91	0.19	0.36
	Best intermediate	265.19	.84	.92	0.19	0.38
	Prototype	431.23	.66	.82	0.22	0.21
Interior-exterior	Exemplar	209.14	.71	.86	0.33	0.14
	Best intermediate	226.02	.62	.82	0.28	0.17
	Prototype	447.88	.15	.39	0.08	0.77
Crisscross	Exemplar	473.29	.75	.87	0.4	0.25
	Best intermediate	406.5	.82	.91	0.42	0.27
	Prototype	1457.8	.01	.12	0.05	0.16

fits. The results of this analysis are depicted in Figure 8A. The vertical axis reflects the number of participants for which a model provided the best fit when comparing the best model with some amount of abstraction with the exemplar model. The color coding reflects whether the difference in deviance was significant (gold) or not (red)

based on a chi-square test. The insets depict the same data as the 3-D plot, now binned based on levels of evidence (left-right: negligible, positive, strong, and very strong) when interpreting the Bayes factor. Figure 8B shows the representation of the model for which the greatest number of participants in a task had a significant

Figure 8. Results of modeling individual observer categorization task choice frequencies using a neural varying abstraction model. (A) Three-dimensional bar histogram of participants based on the best model fit when comparing the exemplar model and the best fitting model with some level of abstraction (including the prototype model). Participants are color coded based on whether the difference in fit between these models was significant (gold) or not (red). Insets plot the same data based on the difference in minimum deviance between the models, as an approximation of the Bayes factor showing evidence in favor of the better fitting model (left-right: negligible, weak, strong, and very strong). Other figure details are the same as in Figure 7. (B) Category representation for the models with the greatest number of significantly superior fits. Other figure details are the same as in Figure 7B.

difference in fits. Again, we will summarize the individual modeling results for each task in turn.

First, for the dimensional task, participant fits were widely distributed across the models. In general, the evidence for nonexemplar models was negligible to positive. Although the exemplar model was numerically not the best fitting for any of the participants, for 9 of 20 participants, the fits were virtually identical (mean difference = 0.35, $SD = 0.47$), whereas for 11 of 20 participants, there was positive evidence for some level of abstraction; for only 5 of 11 were the fits significantly different from the exemplar model (mean difference = 4.89, $SD = 0.72$). For four of five of these participants, the best fitting models involved a single prototype for Category B, the low SF grouping. Interestingly, when looking at the group fits, models with a single subprototype for Category B also did substantially better than models with greater abstraction for Category A. These results suggest that there may be greater use of abstraction in this task than what was revealed by the group level analysis.

Second, for the diagonal task, individual fits were again somewhat widely distributed, with two participants showing negligible evidence in favor of the exemplar model (deviance < 2). For 8 of 19 participants, the best fitting intermediate or prototype model was significantly better than the exemplar model (mean difference = 11.80, $SD = 11.19$). Interestingly, in marked contrast with the group fits, where the prototype model was almost the worst performing model, the prototype model showed the greatest number of significant fits (3/8). This result suggests that averaging across participants may have obscured the extent to which individual participants were relying on more abstract models. Furthermore, again as with the dimensional task, the significantly best fitting models involved more abstraction for low-SF Category B stimuli.

Third, for the interior–exterior task, there was a more clear clustering of participants and 10 of 13 showed significant differences in fits, with most showing strong to very strong evidence in favor of some model representation (mean difference = 15.71, $SD = 11.12$). One cluster involved little abstraction, including 4 of 10 who favored the exemplar representation. The other cluster consisted of participants who tended to favor more abstraction for Category A, the interior grouping, as one might expect from the category structure of the task. Overall, these results are more consistent with the group fits, which clearly favored the exemplar model, but might have obscured the fact that, for six participants, their data were better fit by a model involving some level of abstraction.

Fourth, for the crisscross task, the distribution of fits was the narrowest of all the tasks. For 14 of 18, the best fitting model was significantly different from, and tended to provide strong or very strong evidence against, the exemplar model (mean difference = 10.77, $SD = 5.62$). The vast majority of these participants, 9 of 14, favored a

model representation with one subprototype fusion for Category A, which was the best fitting model for the pooled choice frequencies. Four of 14 participants showing significant fits also seemed to represent subprototypes for Category B stimuli.

Two general trends are worth emphasizing about these individual modeling results. The first is that the group fits underestimate the extent to which observers might be employing categorical representations involving some level of abstraction. This was even true for the interior–exterior task, where a larger proportion of participants seemed to rely on an exemplar representation. The second trend is that, often, this abstraction was for stimuli with low SFs, which required fusions of stimulus representations spanning relatively large divides in the V1 space. This is worth emphasizing, because it speaks against the idea that abstraction might be driven primarily by relative similarity of stimulus representations. Were that the case, then we would expect better fits for models that fused across stimulus orientations. Overall, these individual fit results provide more compelling evidence of a varying role for abstraction in the representation of novel categories.

DISCUSSION

A central focus of psychological research on categorization is the extent to which learning novel categories involves abstracting across differences in stimulus properties. In this study, using VAM, we sought to determine whether observers might utilize distance function applied to neural representations of stimulus dimensions that involve intermediary in their level of abstraction. In so doing, we followed the trend of neuroimaging studies that have moved away from the classic dichotomy between exemplar and prototype models (Mack et al., 2016; Davis et al., 2012). Using a low-dimensional neural space constructed from V1, we compared several VAM representations based on how well they predicted both group and individual observer choice frequencies across four categorization tasks that involved generalizing to novel stimuli.

At the group level, the prototype model provided a poorer fit than the exemplar model across all tasks. For the interior–exterior task, the exemplar model provided substantially better fits to the choice frequencies. For the dimensional and diagonal tasks, several other intermediary representation models provided almost identical fits to the exemplar model. For the crisscross task, an intermediary model with a modest degree of abstraction provided a fit that was significantly better than the exemplar model. At the individual level, there was greater evidence for abstraction across tasks with a good deal of variability between participants. For the dimensional and diagonal tasks, there was a trend toward representations with greater abstraction for low-SF categories, and although many participants favored an exemplar representation for the interior–exterior task (consistent

with the group level fits), an equal proportion favored representations for the interior category with some abstraction. With respect to the crisscross task, the results were highly consistent with the group level results and favored again greater abstraction for low-SF stimuli. Taken together, these results suggest a variable role for abstraction, especially across observers, in category models constructed from neural representations of stimulus dimensions.

Evidence for Abstraction in the Neural Representation of Novel Categories

It is useful to compare our results with those of other studies that have investigated the extent to which abstraction plays a role in the neural representation of novel object categories in visual cortex and other brain regions.

Mack et al. (2013) found that, when using searchlight methods to compare the representations of exemplar and prototype models with neural responses, the exemplar model provided a better match across individuals and was especially correlated with responses in lateral occipital and early visual cortices. Although there are many other methodological differences between their study and ours, there are three worth emphasizing. First, we measured the neural representation of stimulus properties directly and included this in our modeling approach, whereas Mack et al. (2013) modeled observer categorization behavior separately and then compared the model representations with neural responses. Second, by employing VAM, we also tested a larger range of models with intermediate levels of abstraction. Third, in their study, the stimuli varied along four binary feature dimensions (e.g., red/green and triangle/circle), making abstraction psychologically difficult, and had a 4/5 structure that is arguably also biased in favor of exemplar models (Smith & Minda, 2000). Thus, in terms of the MVPA approach, models considered, and stimulus selected, their findings may be as expected, whereas by testing more models and using stimuli that varied along two dimensions, there was a greater possibility for abstraction playing a role in our study.

In identifying some role for abstraction, our results are more in line with neuroimaging studies that have investigated the neural basis of rule-plus-exception learning. In particular, some studies have made use of SUSTAIN, which simulates how observers update category representations across trials (Love et al., 2004). This model also incorporates attentional weighting of feature dimensions, along with a similarity-based clustering process that can be updated to add new clusters to capture surprising class membership. In this model, simple categories defined by a rule or central tendency will result in a prototype representation, but if categories become more complex, then SUSTAIN will store a representation for each item like an exemplar model. Crucially, this

flexibility allows SUSTAIN, like VAM, to generate intermediate representations consisting of potentially multiple subprototypes. Davis et al. (2012) found that SUSTAIN well predicted the neural response of medial temporal lobe activity based on trial-by-trial predictions for item recognition and error correction. Most germane to our results, Mack et al. (2016) had participants learn both a unidimensional task and an XOR (i.e., crisscross) task for their stimuli. They found task-specific differences in attentional weighting of feature dimensions in the hippocampus, with dynamic updating of the region by pFC being greater during early learning phases. Notably, although their modeling approach compares favorably with VAM, these studies also used binary stimulus features (body components of beetles) unlike this study.

One question is the extent to which our results provide evidence against exemplar models as a means of capturing the neural basis of category learning. On this topic, some further points are worth making. First, within the context of comparing categorization models, it is arguable that even showing a tie (e.g., as was the case for many of the group level fits) between exemplar and intermediary models favors the latter, based on a principle of simplicity regarding category representations and decision rules (Peters, Gabbiani, & Koch, 2003). Second, it is also worth considering that we used a small number of stimuli, which varied by much larger steps along each dimension than would be necessary for reliable behavioral stimulus discrimination alone, because we required both neural and behavioral discriminability (and the former is often more difficult to obtain). This is important to emphasize because this requirement likely made our stimuli easy to remember, which can also favor exemplar models (Blair & Homa, 2003). Thus, that we found any evidence favoring some level of abstraction, given these methodological handicaps, is notable. However, clearly, this study is far from the last word on the issue of abstraction and categorical learning in the visual system.

Locating Learned Category Representations in the Visual System

One question is the implication of our results for the neural loci of representations for learned object categories. Our ROI was selected because it is well known to code for both stimulus dimensions under consideration, but it is most plausibly a precursor to whatever region implements the representation of learned categories. This functional division reflects the same distinction in category models between the representation of the stimulus space versus the representation of the learned categories. For this reason, we have been careful to talk of the distance functions applied to the neurally represented stimulus space, not to the category representations themselves, which are likely to be found in higher-level visual areas such as the lateral occipital complex, medial

temporal lobe, or perhaps even the hippocampus (Mack et al., 2013, 2016; Davis et al., 2012).

The distinction between the stimulus space and the category representations is also relevant to interpreting our results with respect to abstraction. For example, one possible concern is that, because V1 codes for orientation and SF, individual stimuli will be represented as quite distinct in this region, which will bias our modeling results in favor of less abstract representations. However, to apply the distance functions of the models at all, it is required that individual stimuli be reliably distinguishable in the neural space; otherwise, there would be no basis with which to assume they are distinctly represented, and model comparison would be impossible. Thus, for our modeling approach and stimulus set, V1 is likely the most appropriate ROI to utilize. Furthermore, the fact that for the individual fits we observed greater abstraction across levels of SF, which were generally greater by a factor of 2/1 to orientation in the V1 neural space, suggests that any overall bias in fitting created solely by our selection of ROI is likely minimal.

Neural Tuning and Psychological Spaces for Stimulus Properties

We measured the discriminability between voxel patterns in V1 based on how stimuli varied along the two dimensions; however, the neural tuning for these properties occurs at the much smaller spatial scale of single cells. There is some correspondence between tuning at the single cell and overall selectivity of individual voxels, which reflect a population response. For example, the classic Gabor wavelet pyramid model of V1 posits a joint code for stimulus position, orientation, and SF (Jones & Palmer, 1987; Daugman, 1985), and the same model has been used to capture selectivity at the voxel level in V1 (Naselaris, Olman, Stansbury, Ugurbil, & Gallant, 2015; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Kay, Naselaris, Prenger, & Gallant, 2008). Plausibly, our ability to carry out run-plus-level decoding is a product of the joint selectivity of V1 voxels for these stimulus properties. However, it is unclear to what extent the selectivity to orientation and SF of individual voxels reflects the underlying neural tuning of a population of single cells, because fMRI only allows the measure of coarse scale maps of neural activity. In particular, because of the subvoxel scale of orientation columns in human V1, it remains unclear what neural signal explains orientation decoding in the region (Carlson, 2014; Freeman, Brouwer, Heeger, & Merriam, 2011; Op de Beeck, 2010; Kamitani & Tong, 2005). Thus, although we were able to construct a 2-D space in which orientation and SF were independent dimensions, it should not be interpreted as a direct reflection of the underlying representational “geometry” (Ritchie, Kaplan, & Klein, 2017; de-Wit, Alexander, Ekroll, & Wagemans, 2016; Naselaris & Kay, 2015).

A Direct Model-based Approach for Investigating Abstraction in the Neural Representation of Novel Categories

A variety of approaches can be taken to relate formal models of categorization to neural signals (Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017; Forstmann, Wagenmakers, Eichele, Brown, & Serences, 2011). We focused on constructing models from a neural, as opposed to a psychological, space for our stimulus dimensions (Davis & Poldrack, 2014; Op de Beeck et al., 2008), which allowed for a direct model comparison in line with the sorts of analyses traditionally used to compare formal models of categorization (Churchland & Kiani, 2016). In this regard, we aimed for an analysis that was both biologically and psychologically plausible (Ritchie & Carlson, 2016). However, our direct modeling approach also required making explicit algorithmic and parametric assumptions to use neural distances to predict observer responses, and these assumptions may at best loosely approximate neural processing. Still, we believe VAM provides a useful tool for comparing different possible category representations, even if other formal models might more accurately capture the neural process of category learning. For example, distance-to-bound models of categorization apply a distance function between a decision boundary and the position of a stimulus representation in psychological space (Ashby & Townsend, 1986) and are virtually identical in architecture to some linear classifiers (Ritchie & Carlson, 2016). These models offer a very different approach to category learning than both prototype and exemplar models and may more closely approximate the process of linear readout applied to neural representations of learned categories, although it is worth noting that distance-to-bound models can also be formally connected to both prototype and exemplar models (Ashby & Maddox, 1993).

Limitations and Future Directions

Our experimental design mimicked classic behavioral studies on category learning in which separate groups of participants perform the similarity and categorization tasks. A benefit of this approach was that it allowed us to test models of choice frequencies for many different tasks but precluded detecting neural signatures of category learning. Previous neuroimaging studies that have investigated abstraction have not had this limitation. For example, it has been found that attentionally weighted models better capture the neural representations in regions such as the lateral occipital cortex (Mack et al., 2013) and the hippocampus (Mack et al., 2016). Similarly, several studies have shown effects of perceptual learning, which can be thought of as a form of local attentional weighting (Aha & Goldstone, 1992), in early visual cortex (Jehee, Ling, Swisher, van Bergen, & Tong, 2012; Shibata

& Watanabe, 2011). This shortcoming points to a number of possible directions for future research.

First, future research might try to jointly isolate distinct regions recruited during the learning and application of category representations and the relationship between these regions based on attentional weighting and perceptual learning (Mack et al., 2013, 2016; Jehee et al., 2012; Shibata & Watanabe, 2011). Second, other studies looking at select intermediary models have investigated how category representations are built up on a trial-by-trial basis and their relation to neural responses (Mack et al., 2016; Davis et al., 2012). There is the potential to take a similar approach using VAM to investigate how neural representations of stimulus dimensions are influenced by task demands. Third, although individual models were tested separately at both the group and individual levels, this could also be done jointly using a hierarchical Bayesian approach (Bartlema et al., 2014; Lee & Wagenmakers, 2014). Any of these directions could be pursued to further investigate how the distribution of fits across multiple models, which vary in levels of abstraction, might differ based on task when observers categorize stimuli in the scanner.

Summary

Abstraction is a basic principle for differentiating representations of learned categories. Ours is the first study to construct a class of categorization models, varying in abstraction, directly from a low-dimensional neural space representing the stimulus dimensions. Our results, which favor a variable role for abstraction in category models constructed from a neural space, illustrate the value of an underexplored direction for marrying formal models and neuroimaging findings.

Acknowledgments

This work was supported by the European Research Council (ERC-2011-StG-284101), a federal research action grant (IUAP-P7/11), and a Hercules grant (ZW11_10) to H. O. P. This project has received funding from the FWO and European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant agreement no. 665501, via an FWO [PEGASUS]² Marie Skłodowska-Curie fellowship (12T9217N) to J. B. R.

Reprint requests should be sent to J. Brendan Ritchie, Psychologisch Instituut, Tiensestraat 102, Box 3714, 3000 Leuven, Belgium, or via e-mail: brendan.ritchie@kuleuven.be.

REFERENCES

- Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 534–539). Citeseer.
- Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 598–612.
- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37, 372–400.
- Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154–179.
- Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology*, 63, 516.
- Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, 59, 132–150.
- Benson, N. C., Butt, O. H., Brainard, D. H., & Aguirre, G. K. (2014). Correction of distortion in flattened representations of the cortical surface allows prediction of V1–V3 functional organization from anatomy. *PLoS Computational Biology*, 10, e1003538.
- Benson, N. C., Butt, O. H., Datta, R., Radoueva, P. D., Brainard, D. H., & Aguirre, G. K. (2012). The retinotopic organization of striate cortex is well predicted by surface topology. *Current Biology*, 22, 2081–2085.
- Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5-4 categories and the category advantage. *Memory & Cognition*, 31, 1293–1301.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.
- Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, 118, 2–16.
- Carlson, T. A. (2014). Orientation decoding in human visual cortex: New insights from an unbiased perspective. *Journal of Neuroscience*, 34, 8373–8383.
- Chen, Y., Wiesel, A., Eldar, Y. C., & Hero, A. O. (2010). Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58, 5016–5029.
- Churchland, A. K., & Kiani, R. (2016). Three challenges for connecting model to mechanism in decision-making. *Current Opinion in Behavioral Sciences*, 11, 74–80.
- Coutanche, M. N., & Thompson-Schill, S. L. (2012). The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. *Neuroimage*, 61, 1113–1119.
- Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2, 1160.
- Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, 22, 260–273.
- Davis, T., & Poldrack, R. A. (2014). Quantifying the internal structure of categories using a neural typicality measure. *Cerebral Cortex*, 24, 1720–1737.
- de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin and Review*, 23, 1415–1428.
- Fischl, B. (2012). FreeSurfer. *Neuroimage*, 62, 774–781.
- Forstmann, B. U., Wagenmakers, E. J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? *Trends in Cognitive Sciences*, 15, 272–279.
- Freeman, J., Brouwer, G. J., Heeger, D. J., & Merriam, E. P. (2011). Orientation decoding depends on maps, not columns. *Journal of Neuroscience*, 31, 4792–4804.
- Gauthier, I., & Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? *Annual Review of Vision Science*, 2, 377–396.

- Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15, 536–548.
- Harvey, B. M., & Dumoulin, S. O. (2011). The relationship between cortical magnification factor and population receptive field size in human visual cortex: Constancies in cortical architecture. *Journal of Neuroscience*, 31, 13604–13612.
- Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37, 435–456.
- Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8, 686–691.
- Henriksson, L., Nurminen, L., Hyvarinen, A., & Vanni, S. (2008). Spatial frequency tuning in human retinotopic visual areas. *Journal of Vision*, 8, 5.
- Jehee, J. F. M., Ling, S., Swisher, J. D., van Bergen, R. S., & Tong, F. (2012). Perceptual learning selectively refines orientation representations in early visual cortex. *Journal of Neuroscience*, 32, 16747–16753.
- Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58, 1233–1258.
- Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8, 679–685.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *The American Statistical Association*, 90, 773–795.
- Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452, 352–355.
- Kriegeskorte, N. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 1–28.
- Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17, 401–412.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27.
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian modeling for cognitive science: A practical course*. Cambridge, UK: Cambridge University Press.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111, 309–332.
- Luce, R. D. (1959). *Individual choice behavior*. Oxford, UK: John Wiley.
- 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, U.S.A.*, 113, 13203–13208.
- Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, 23, 2023–2027.
- Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MPPA using explicit models of representation. *Trends in Cognitive Sciences*, 19, 551–554.
- Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage*, 105, 215–228.
- Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63, 902–915.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 87–108.
- Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, 45, 279–290.
- Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 924–940.
- Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMPPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, 10, 1–27.
- Op de Beeck, H. P. (2010). Probing the mysterious underpinnings of multi-voxel fMRI analyses. *Neuroimage*, 50, 567–571.
- Op de Beeck, H. P., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, 132, 491–511.
- Op de Beeck, H. P., Wagemans, J., & Vogels, R. (2008). The representation of perceived shape similarity and its role for category learning in monkeys: A modeling study. *Vision Research*, 48, 598–610.
- Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, 5, 291–303.
- Peters, R. J., Gabbiani, F., & Koch, C. (2003). Human visual object categorization can be described by models with low memory capacity. *Vision Research*, 43, 2265–2280.
- Plummer, M. (2003). *A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.3406>.
- Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 353–363.
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3, 382–407.
- Ritchie, J. B., Bracci, S., & Op de Beeck, H. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *Neuroimage*, 148, 197–200.
- Ritchie, J. B., & Carlson, T. A. (2016). Neural decoding and “inner” psychophysics: A distance-to-bound approach for linking mind, brain, and behavior. *Frontiers in Neuroscience*, 10, 1–8.
- Ritchie, J. B., Kaplan, D. M., & Klein, C. (2017). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*. <https://doi.org/10.1093/bjps/axx023>.
- Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1, 54–87.
- Shibata, K., & Watanabe, T. (2011). Perceptual learning incepted by decoded fMRI neurofeedback. *Science*, 334, 1413–1416.
- Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1411–1436.

- Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 3–27.
- Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76, 65–79.
- Vanpaemel, W. (2009). BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods*, 41, 1111–1120.
- Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, 15, 732–749.
- Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, 42, 421–437.
- Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. *Proceedings of the Annual Conference of the Cognitive Science Society*, 27, 2277–2282. Retrieved from <https://lirias.kuleuven.be/bitstream/123456789/150831/1/vanpaemelso2005.pdf>.
- Verbeemen, T., Vanpaemel, W., Pattyn, S., Storms, G., & Verguts, T. (2007). Beyond exemplars and prototypes as memory representations of natural concepts: A clustering approach. *Journal of Memory and Language*, 56, 537–554.
- Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, 137, 188–200.
- Wardle, S. G., Ritchie, J. B., Seymour, K., & Carlson, T. A. (2017). Edge-related activity is not necessary to explain orientation decoding in human visual cortex. *The Journal of Neuroscience*, 37, 1187–1196.

Uncorrected Proof