**Title:** Category learning selectively enhances representations of boundary-adjacent exemplars in early visual cortex.

**Abbreviated title:** Category learning in early visual cortex

**Authors:** Sean R. O'Bryan, Shinyoung Jung, Anto J. Mohan, & Miranda Scolari

**Affiliations:** Department of Psychological Sciences, Texas Tech University, 79409

**Corresponding author:** Sean O'Bryan (sean_obryan@brown.edu)

**Number of pages:** 41

**Number of figures:** 5

**Number of words:**

1. **Abstract:** 248
2. **Introduction:** 649
3. **Discussion:** 1375

**Conflict of interest statement:** The authors declare no competing financial interests.

**Acknowledgments:**

S.R.O.'s primary contributions to this work were made at Texas Tech University. He is currently affiliated with the Department of Cognitive, Linguistic & Psychological Sciences, Brown University, 02912.

We thank Thomas C. Sprague for helpful discussions on this project.

**Abstract**

Category learning and visual perception are fundamentally interactive processes, such that successful categorization often depends on the ability to make fine visual discriminations between stimuli that vary on continuously valued dimensions. Research suggests that category learning can improve perceptual discrimination along the stimulus dimensions that predict category membership, and that these perceptual enhancements are a byproduct of functional plasticity in the visual system. However, the precise mechanisms underlying learning-dependent sensory modulation in categorization are not well understood. We hypothesized that category learning leads to a representational sharpening of underlying sensory populations tuned to values at or near the category boundary. Furthermore, such sharpening should occur largely during active learning of new categories. These hypotheses were tested using fMRI and a theoretically constrained model of vision to quantify changes in the shape of orientation representations while human adult subjects learned to categorize physically identical stimuli based on either an orientation rule (N = 12) or an orthogonal spatial frequency rule (N = 13). Consistent with our predictions, modeling results revealed relatively enhanced reconstructed representations of stimulus orientation in visual cortex (V1–V3) only for orientation rule learners. Moreover, these reconstructed representations varied as a function of distance from the category boundary, such that representations for challenging stimuli near the boundary were significantly sharper than those for stimuli at the category centers. These results support an efficient model of plasticity wherein only the sensory populations tuned to the most behaviorally relevant regions of feature space are enhanced during category learning.

## Significance Statement

Poisonous or edible? Friend or foe? Quickly grouping objects into appropriate categories is critical to our survival. Many category decisions are supported by the presence of one or more defining features – for example, the shape and color of a banana can easily distinguish it from other fruits at the store. Other decisions require highly precise perceptual representations – which exact shade of yellow determines whether a banana is ripe? We tested the hypothesis that ongoing learning of new visual categories leads to more precise sensory representations, especially where precision is likely to improve categorization performance. Our results bore this out: active category learning can lead to rapid and specific improvements in the way early visual cortex represents relevant features.

Category learning enables us to predict the behavioral relevance of novel stimuli. In the visual domain, this is made possible by selectively attending to the specific features that lead to successful categorization. For example, noting whether an organism has wings is useful for distinguishing birds from mammals, but uninformative when classifying bird species. Instead of discrete features, bird watchers are better served by attending to continuous dimensions such as color or texture. Learning to categorize such stimuli can lead to improved perception of subtle differences across relevant dimensions, especially for physically similar stimuli that nonetheless belong to distinct categories (Rosch, et al., 1976; Curby & Gauthier, 2010; Diamond & Carey, 1986; Tarr & Gauthier, 2000; Seger et al., 2015; Hamm & McMullen, 1998; Jolicoeur et al., 1984; Zeithamova & Maddox, 2007).

Dimensional relevancy is a likely catalyst for this improved perceptual sensitivity. For instance, categorizing size- and brightness-varying objects by size makes small size differences easier to distinguish, but not brightness differences (Goldstone, 1994). This may be due to perceptual stretching along the relevant dimension, where small feature value differences become exaggerated (Goldstone & Steyvers, 2001; Folstein et al. 2013; Folstein et al., 2015). Neuroimaging and single-unit recording studies support the hypothesis that category learning leads to warped neural representations of relevant exemplars (Sigala & Logothetis, 2002; O'Bryan et al., 2018a, 2018b; but see Jiang et al., 2007), and such neural plasticity may directly support perceptual discrimination (Folstein et al., 2012; Folstein et al., 2013).

Dimension wide perceptual stretching can account for a broad range of results (Nosofsky, 1986). Nonetheless, open arguments suggest category learning should produce localized enhancement for a subset of features along an attended dimension (sometimes termed *categorical perception*). Perceptual noise leads to particularly high classification error rates near

category boundaries (Aha & Goldstone, 1992; Maddox & Ashby, 1993), and as such, precise perceptual representations for these exemplars may be uniquely crucial. If so, classifying visually similar between-category exemplars should lead to enhanced neural representations at or around the boundary – especially during learning when internal boundaries are inherently noisy.

The behavioral evidence for such localized representational enhancement effects have been mixed (Juárez et al., 2019; Folstein et al., 2014; Van Gulick & Gauthier, 2014), where most studies search for demonstrations of persistent perceptual improvements outside of active categorization. However, localized representational enhancement is consistent with the known neurobiology of feature-based selective attention. When nonhuman primates attend to specific feature values (e.g., red), sensory neurons tuned to the most task-informative values exhibit elevated firing rates, whereas responses from neurons tuned to uninformative values (e.g., blue) within the same feature space are often suppressed (Sigala & Logothetis, 2002; Martinez-Trujillo & Treue, 2004; Yang & Manusell, 2004), leading to enhanced representations of relevant sensory input (Ling et al., 2009). Importantly, the perceptual learning literature indicates this representational enhancement is task-dependent, especially in early visual cortex (Byers & Serences, 2014).

Most visual categorization studies have focused on parietal, prefrontal, and extrastriate regions with the expectation that they are uniquely sensitive to learning effects (Freedman & Assad, 2016; Uyar et al., 2016). Few studies have examined the possible downstream effects of category learning on retinotopically organized regions of visual cortex, with the recent exception of Ester et al. (2020). Despite relatively sparse research, there is ample evidence to suggest that V1 may play an integral role during category learning, analogous to its role in perceptual learning.

We address this question using fMRI and an encoding model to reconstruct orientation representations within early visual cortex while subjects actively learn to categorize grating stimuli based on an orientation (line angle) rule or an orthogonal spatial frequency (line width) rule. We predicted orientation representations should be enhanced among orientation learners to optimally support boundary acquisition and minimize prediction error during learning. Furthermore, these sensory modulations should be most pronounced for exemplars that border subjects' assigned category boundaries, consistent with an efficient model of plasticity.

## Methods

### *Subjects*

Twenty-six healthy adult human subjects (age range: 18 – 32 years; 13 females, 12 males, and 0 nonbinary) with normal or corrected-to-normal vision were recruited from the Texas Tech University community. Data from one subject was removed due to excessive movement in the scanner, which resulted in considerable loss of visual cortex coverage. All subjects provided written informed consent before participating in accordance with the Declaration of Helsinki. Subjects were paid $20/hr for the fMRI scanning sessions, and $10/hr for behavioral training completed outside of the scanner. This study was approved by the Texas Tech University IRB.

### *Materials*

Visual stimuli were rendered using MATLAB (v.9.1, MathWorks) and presented via Psychophysics Toolbox (v.3.3; Kleiner et al., 2007) on a desktop PC running Windows 10. For a pre-scan training session, stimuli were displayed on a 1920 x 1080 pixel resolution BenQ XL2430T monitor measuring 58 cm wide and set to a 100 Hz refresh rate. During all fMRI

scans, stimuli were presented on a 1024x768 resolution projection screen measuring 19 cm wide and at a 60 Hz refresh rate.

*Categorization Task*

The primary goal of this experiment is to characterize modulations in orientation-selective population responses while subjects actively learn categories, where orientation is either a category-relevant or irrelevant stimulus dimension. To accomplish this goal, subjects learned to classify grating stimuli into one of two categories via trial and error. Subjects were assigned to one of two experimental conditions based on their subject number: categorization based on either an orientation rule (N = 12) or a spatial frequency rule (N = 13). This group sample size was determined based on related studies obtaining medium to large within-group effects sizes with samples ranging between 8-13 (Scolari et al., 2012; Byer & Serences, 2014; Ester et al., 2020). Assignment to these conditions was performed pseudo-randomly (based on subject number) to ensure an approximately equal number of subjects in each group.

Subjects were not aware of the rule they would learn prior to beginning the categorization task. However, they were informed that the categorization rule may be based on either the orientation or spatial frequency dimensions of the gratings. Critically, all subjects encountered an identical stimulus set over the course of the experiment regardless of their assigned categorization rule; the task differed between subjects only with respect to the categories to which each stimulus belonged.

Procedurally, each trial of the categorization task began with a 3 s grating stimulus. Gratings were presented centrally on a middle gray background with a radius of 8 degrees of visual angle and flickered at a rate of 5 Hz to drive responses in early visual cortex. During both

stimulus presentation and inter-stimulus intervals (ISIs), subjects were instructed to maintain fixation on a black point in the center of the screen. Fixation was monitored in real time by the experimenter via an MRI-compatible eye tracker (Eyelink 1000 Plus; SR Research, Ontario, Canada) to ensure that the retinotopic location of stimuli was consistent both within and between subjects across task conditions.

Subjects responded with a button press corresponding to "Category A" or "Category B" during the 3 s stimulus presentation period. During the last 1 s of the trial, feedback was administered via a color change at central fixation (green and red for correct and incorrect, respectively) while the grating stimulus remained on the screen. Following the 3 s combined stimulus presentation, response, and feedback window, the grating was removed from the screen and subjects encountered a fixation-only inter-stimulus interval (ISI). The duration of each ISI was pseudo-randomly jittered with a mean of 4 s and drawn from a distribution ranging between 2 – 6 s in 500 ms steps (resulting in 9 possible ISI durations encountered equally often during each scanning run). Subjects completed 6 categorization scanning runs of 54 trials each, with a run time of 6 min 20 s.

The exemplars encountered during the experiment varied on the two critical dimensions. Each exemplar took on one of 18 possible values in orientation space, ranging from 5° to 175° in 10° steps. Similarly, exemplars expressed one of 18 possible values in spatial frequency space, ranging from 0.44 cycles/degree to 1.25 cycles/degree in .045 cycle/degree steps. The values for each dimension were randomized throughout the experiment.

For all subjects assigned to learn the spatial frequency rule, the category boundary was defined as the midpoint of the constrained spatial frequency space, with the 9 highest spatial frequencies belonging to Category A and the 9 lowest spatial frequencies belonging to Category

B. For subjects assigned to learn an orientation rule, one of four possible category boundary pairs was assigned based on subject number (20°/110°, 40°/130°, 60°/150°, and 80°/170°). In 180° orientation space, boundary pairs are required because orientation space is circular (Fig. 1).

Within the week prior to their scheduled fMRI scans, subjects attended a brief (< 30 minutes) training session outside of the scanner where they completed two practice blocks of a categorization task. The task employed the same stimuli and response mappings used in the primary fMRI experiment. Critically, however, the categorization rule for these practice blocks was identical across all participants, using a 45/135 degree orientation boundary pair that was not assigned to any subjects for the scanning session. Subjects were told that the categorization rule could be based on either the spatial frequency (line width) or the orientation of the gratings, but were not explicitly informed to which rule they were assigned. The rationale for this brief practice session was to sufficiently familiarize participants with the task procedure and stimuli, and ultimately was expected to support more rapid learning when the categorization task was completed in the scanning environment. On the day of the scanning session, subjects were reminded that they would encounter a new, random rule defined by either the spatial frequency or orientation of the gratings.

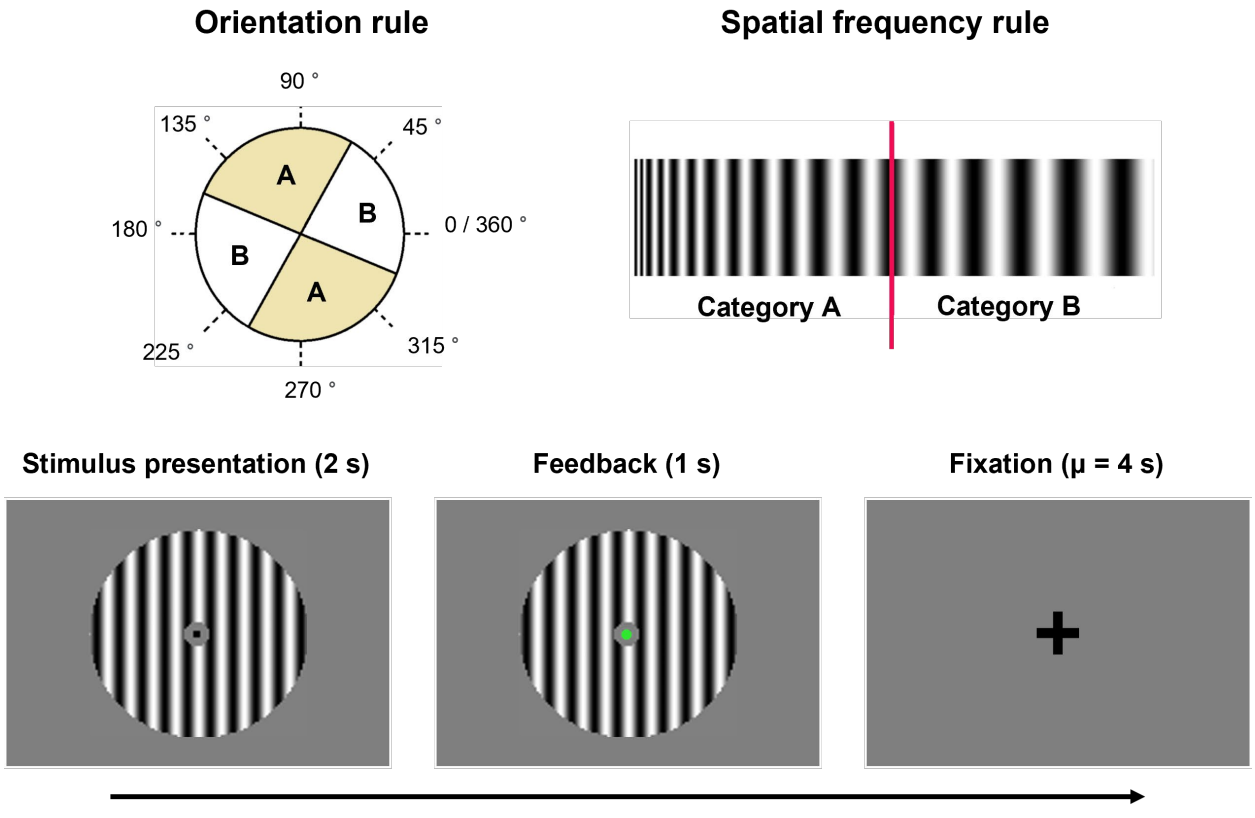


**Figure 1.** Experimental design. For the primary categorization task, subjects learned to categorize grating stimuli according to either an orientation rule based on one of four possible boundary pairs (top-left; 60°/150° boundary pair depicted) or a spatial frequency rule based on a midpoint boundary (top-right). Stimuli and the time course for an example trial are depicted in the bottom row. Note that the trial structure was identical for the contrast discrimination task.

*Orthogonal Contrast Discrimination Task*

To allow for tightly controlled within-subjects comparisons, subjects completed 6 scanning runs of an orthogonal contrast discrimination task made up of the same flickering stimuli used in the categorization task. Here, subjects were required to discriminate between slight increases and decreases in grating contrast. We reasoned that discriminating contrast changes would provide a strong control condition, because this requires that subjects attend to the grating to successfully complete the task (thus matching the presumed spatial extent of attention across tasks).

Once in each trial, the contrast of the grating either decreased or increased for 100 ms (within a single flicker cycle). Subjects were instructed to press a button with their index finger to indicate a perceived decrease in contrast, and with their middle finger to indicate a perceived increase in contrast. As with the categorization task, feedback was administered in the form of a red or green fixation point appearing on the screen for the final 1 s of the 3 s stimulus presentation window. Each trial was separated by a jittered ISI with the same parameters used in the categorization task described above.

To allow enough time for subjects to respond and receive feedback during the 3 s stimulus presentation window, the brief contrast changes were applied at pseudo-random intervals within the first 1.5 s of stimulus onset. For the first run of the contrast task, the magnitude of contrast changes (both increases and decreases) started at a default of 20%. After the first run, task difficulty was manually titrated by the experimenter on a run-by-run basis to approximately match expected performance in the categorization task by increasing or decreasing the magnitude of contrast change for each run in 5-10% increments.

Importantly, all contrast scans were run first to ensure subjects did not engage in orientation or spatial frequency categorization during the task. The same contrast changes were then implemented on a scan-by-scan basis during the categorization task to perfectly equate all stimulus properties across the two study phases, but these changes were irrelevant during categorization.

*Retinotopic Mapping*

All subjects recruited for the study completed a separate, standard retinotopic mapping scan. This procedure is used to identify and map early visual cortical areas (V1, V2, and V3)

unique to each subject. The scans required passive fixation on a rotating checkerboard stimulus, subtending 60° of visual angle and flickering at a rate of 8 Hz (Engel et al., 1994; Sereno et al., 1995; Swisher et al., 2007; Arcaro et al., 2009). To ensure that subjects were attentive throughout the scan, they were instructed to press a button with their right index finger when they detected a gray segment that periodically appeared in the stimulus display. The functional datasets were later projected onto an inflated representation of cortex for each subject to demarcate the functional borders between visual areas V1v, V1d, V2v, V2d, V3v, and V3d.

*fMRI Data Acquisition and Preprocessing*

Imaging data were acquired on a 3.0 T Siemens Skyra MRI scanner at the Texas Tech Neuroimaging Institute. MPRAGE anatomical scans (two collected during the retinotopy scan session; one collected during the experimental scan session) provided high-resolution structural images of the whole brain in the sagittal plane for each participant (TR = 2.5 s; TE = 1.7 ms; $\theta$ = 7°; slice thickness = 1 mm, slices = 172). Functional images were acquired using a single-shot T2*-weighted gradient echo EPI sequence (TR = 2 s; TE = 40 ms; $\theta$ = 72°; FoV= 256 mm; matrix = 128 x 128 mm; number of axial slices = 25, voxel size = 2x2x3 mm with 0.5 mm gap), and slices were oriented to cover the full extent of the occipital lobe.

Data preprocessing was carried out using AFNI and SUMA with custom time series analysis routines for slice-time correction, between- and within-scan motion correction, and high-pass temporal filtering (3 cycles/run). Voxel time series were normalized ($z$-scored) within run to correct for differences in mean signal intensity across voxels, and trial-level activation in each voxel was demeaned to ensure that evidence of orientation selectivity can be attributed to

the activation patterns in orientation-selective cortex as opposed to mean changes in the BOLD response across voxels that may be evoked by different orientations.

*fMRI Analysis*

For the primary categorization task and orthogonal contrast discrimination task, 3 s trial-level BOLD responses were estimated using block regressors in AFNI's 3Ddeconvolve program. Estimates for the amplitude of the BOLD response on each trial served as input for the inverted encoding model described below to generate estimates for the reconstructed orientation representations associated with each task condition. Data were spatially smoothed using a 4 mm FWHM Gaussian kernel. Prior to training the inverted encoding model, a voxel selection procedure was performed to identify subsets of voxels in V1, V2, and V3 that best distinguished between differing orientation values. To do so, F-values for a one-factor ANOVA with orientation as the single factor were computed for all voxels in each independent training set, where the top 25% of orientation-selective voxels were then used for the given model training and testing iteration.

*Inverted Encoding Model Analyses*

The BOLD responses observed for identified orientation-selective voxels represent the summed activity of many individual orientation-selective neurons. Although the neurons contributing to the BOLD signal in each voxel may be associated with different underlying orientation preferences, research suggests that voxels in early visual cortex exhibit small but reliable biases in orientation sensitivity (e.g., Kamitani & Tong, 2005; Serences et al., 2009; Jia et al., 2011). These consistent biases can be leveraged to make quantitative predictions about how representations of stimulus orientation have changed across visual cortex as a result of task

demands or voluntary attention. This approach was adopted under the premise that learning categories defined by an orientation rule may lead to shifts in the amplitude, slope, and/or bandwidth of population-wide response functions (Byers & Serences, 2014), particularly for challenging stimuli falling near the category boundaries.

To generate these predictions, we employed an inverted encoding model (Brouwer & Heeger, 2009; 2011; Scolari et al., 2012; Byers & Serences, 2014; Sprague et al., 2018; Ester et al., 2020). Encoding models make theoretically motivated assumptions about how relevant features are represented in the brain. When subjects encounter visual features represented in this model, the resulting BOLD response can be used to weight voxels according to the similarity between their true response and the theoretical response for each feature. Finally, the model is "inverted," such that the voxel weights associated with each feature are used to reconstruct channel response functions (CRFs) using independent task data.

Functions used for the theoretical basis set in the model were based on well-established single-unit tuning functions in V1 associated with orientation perception. Specifically, the model assumes each orientation tuning function to be half-sinusoidal in shape and raised to the $9^{th}$ power, where the half-bandwidth of orientation selective neurons spans 20° of orientation space. The model requires a minimum number of evenly spaced functions such that the entire 180° space is covered, and the maximum number of functions should not exceed the number of unique features presented in order to avoid overfitting. To both satisfy these criteria and to maintain consistency with previous studies (Scolari et al., 2012), we used a basis set of 10 evenly distributed orientation functions in the current experiment.

The a priori model parameters described above were incorporated into an encoding model first described by Brouwer and Heeger (2009; 2011) with the goal of reconstructing orientation

representations associated with different task conditions. Formally, the model requires input parameters for the number of voxels selected ($m$), the number of trials in the training or testing datasets ($n$), and the number of pre-defined orientation channels ($k$, where $k = 10$ for the current study). $B_1$ and $B_2$ represent $m \times n$ matrices used to denote the training and testing datasets. The datasets were defined using a leave-one-out approach where the model was trained using data from 10 total scanning runs-- five from the contrast discrimination task and five from the categorization task-- with one run from each task used separately as the test dataset for each iteration of the model. The training data ($B_1$) was mapped on to the full rank matrix of hypothetical channel outputs ($C_1$, $k \times n$) using a weight matrix ($W$, $m \times k$) estimated from the training data using a GLM:

$$B_1 = WC_1, \quad (1)$$

Where the ordinary least-squares estimate of $W$ is computed as follows:

$$W_{\text{fitted}} = B_1C_1'(C_1C_1')^{-1} \quad (2)$$

The channel responses $C_2$ for each trial were then estimated for the test data $B_2$ by applying the fitted weights from equation 2:

$$C_{2\ \text{fitted}} = (W_{\text{fitted}}'\ W_{\text{fitted}})^{-1}\ W_{\text{fitted}}'B_2 \quad (3)$$

Channel responses corresponding to each of the 10 specified orientation channels were then circularly shifted for each trial and projected into 180 degree orientation space, such that the orientation of each presented stimulus is depicted at the center of the resulting CRF. After iteratively performing the leave-one-out cross validation approach with each pair of scanning runs as the test datasets, the CRFs estimated for each scanning run were averaged across all runs for each task condition (e.g. orientation categorization; spatial frequency categorization; contrast discrimination) for statistical comparison. These CRFs were then binned according to the

distance between the presented stimulus and subjects' assigned orientation boundaries to test the hypotheses of a graded representational enhancement as stimuli approached the category boundaries in the orientation rule condition.

Finally, each subject's averaged CRF was fit with the following exponential cosine function (Byers & Serences, 2014; Ester et al., 2020):

$$f(x) = \alpha\left(e^{k(\cos(\mu - x) - 1)}\right) + \beta$$

where $x$ corresponds to the channel responses, $\alpha$ is the vertical scaling (restricted to a range of 0 to 3), $k$ is the concentration (which determines the width; restricted to a range of 0.125 to 100), $\mu$ is the function's center (restricted to a range of 0 to pi), and $\beta$ is the baseline (restricted to a range of -3 to 3). These model fits were then used to quantify the shape of the reconstructed representations. To test our hypotheses about whether learning to categorize oriented gratings leads to stronger, sharper, and/or more precise representations of orientation values, we report amplitude (the difference between the maximum and minimum estimated values); slope; bandwidth (the inverse of concentration); and center shifts (the absolute difference between the presented orientation and the estimated center).

*Inverted Encoding Model Predictions*

Our task design afforded us the opportunity to test for possible changes in orientation representations both within and between subjects. First, the orthogonal contrast detection task served as a stimulus-matched comparison condition to determine if learning to categorize stimuli based on orientation enhances the neural representation of behaviorally relevant feature values in visual cortex. We anticipated that the reconstructed representations of stimulus orientation should be relatively enhanced during the categorization task compared to the contrast

discrimination task for subjects assigned an orientation rule. This enhancement could take the form of higher amplitudes, steeper slopes, and/or narrower bandwidths. Such enhanced representations may be most beneficial when they are centered on or near the presented orientation value, especially during early learning when participants are engaged in active exploration of the category space. Thus, we might also expect the estimated function centers to be closer to the presented orientation value during orientation categorization compared to contrast discrimination. Conversely, we expected no differences in any of these measures between the categorization and contrast tasks among the spatial frequency group.

We furthermore predicted that subjects learning to categorize stimuli based on orientation would exhibit enhanced orientation representations that are specifically relevant to categorization decisions. In particular, we expected representational enhancement to be most prominent for stimuli near subjects' assigned category boundaries in the orientation group compared to exemplars at the center of each category.

Offline, we randomly applied one of the four orientation boundary pairs to each of the spatial frequency learners' data to accommodate between-subject comparisons of orientation representations for near and far boundary trials. Importantly, we used the same boundary pairs that were assigned to the orientation group, so that the boundaries were fully matched between groups. For the spatial frequency group, we expected the shape of the resulting CRFs to be uniform, as no significant representational differences should occur for stimulus values that are near or far from arbitrarily assigned orientation boundaries.

It is possible that boundary-specific enhancement effects emerge at a specific stage of learning. For example, enhanced sensory representations of stimuli may only be beneficial during early learning when many errors are committed around the category boundary.

Alternatively, such effects may instead emerge only after an adequately high level of performance is achieved in the task. To address these possibilities, we divided the data into early (blocks 1 and 2) and late (blocks 5 and 6) learning stages to test whether learning duration differentially modulates the shape of reconstructed representations for near- and far-boundary stimuli.

## Results

### *Learning Performance*

To ensure the spatial frequency categorization task was an appropriate control for orientation categorization, we first compared mean task accuracy and asymptotic learning between both groups. Mean categorization accuracy was well above chance among both the orientation ($M$ = 83.0%, $SD$ = 11.8%) and spatial frequency (85.1%, $SD$ = 4.1%) groups across the 6 categorization blocks. Critically, all subjects learned their respective category rules as indicated by accuracy on the last two learning blocks (orientation: range = 62.0% - 96.3%; spatial frequency: range = 75.9% - 92.6%), including the worst-performing subject whose accuracy remained significantly above chance, t (107) = 2.57, $p$ = .006.

To ensure that category learning was well matched between groups, we used a linear mixed model with factors for learning block, categorization group, and their interaction. The model revealed a significant main effect of block on accuracy, $F$ (5, 115) = 6.94, $p$ < .001, while neither the main effect of group, $F$ (1, 23) = .35, $p$ = .56, nor block × group interaction, $F$ (5, 115) = .60, $p$ = .70, were significant. Taken together, these results suggest that the accuracy of both groups improved significantly over the course of the 6 learning blocks, and that these

improvements did not differ between categorization rules (see Fig. 2a). Thus, the spatial frequency rule served as an appropriate control condition to the orientation rule.

Secondary within-subject comparisons were carried out to assess performance on the stimulus-matched perceptual discrimination task relative to categorization. Mean accuracy for contrast discrimination was somewhat lower than that observed for the categorization task in both the orientation ($M$ = 74.9%, $SD$ = 15.1%) and spatial frequency groups ($M$ = 79.5%, $SD$ = 6.2%). Linear mixed models with accuracy as the outcome variable and factors for task (contrast vs. categorization), block, and their interaction revealed a significant main effect of task for the orientation group, $F$ (1, 11) = 7.42, $p$ = .02, with categorization accuracy being higher than contrast discrimination accuracy on average (Fig 2b). For the spatial frequency group, a significant task × block interaction was observed, $F$ (5, 120) = 3.54, $p$ = .005, such that relative differences in accuracy were larger for spatial frequency subjects on the categorization task relative to the contrast task for early blocks, but not late blocks (Fig 2c). Importantly, mean performance on the contrast discrimination task did not significantly differ between the orientation and spatial frequency subjects, t (23) = -.99, $p$ = .33, $d$ = 0.43. These results suggest that the orthogonal contrast discrimination task was slightly more difficult than the subsequent categorization tasks completed by both groups, but critically, that these differences were largely equated between the experimental groups.

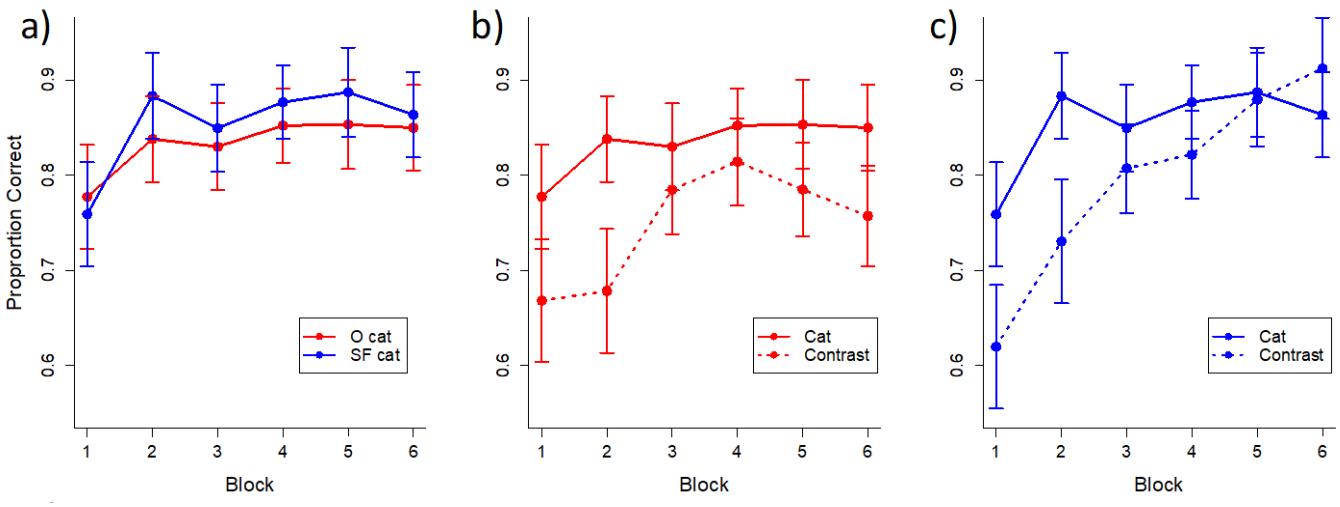


**Figure 2.** Learning curves. Red points correspond to the orientation group, and blue points correspond to the spatial frequency group. Solid lines indicate accuracy (proportion correct) during the categorization task, whereas dotted lines indicate accuracy during the contrast task. a) Orientation versus spatial frequency categorization. b) Orientation categorization versus contrast discrimination. c) Spatial frequency categorization versus contrast discrimination. Error bars reflect standard error of the mean (SEM).

*Channel response functions for categorization vs. contrast discrimination tasks*

We predicted relative increases in amplitude and slope, as well as possible decreases in bandwidth and center shift of orientation CRFs, when comparing the orientation categorization task to the orthogonal, physically matched contrast discrimination task. This prediction is based on the broader theory that fine perceptual discriminations between continuously valued stimuli may be supported by stronger (e.g., higher amplitude) and/or more specific (e.g., steeper slopes) neural representations of task-relevant features in the sensory populations responsible for their perception (Byers & Serences, 2014; Scolari et al., 2012). We were particularly interested in testing the interaction between task (contrast vs. categorization) and category learning condition (orientation vs. spatial frequency), as evidence of representational enhancement.

Linear mixed models including factors for categorization condition (orientation and spatial frequency), task phase (contrast discrimination and categorization), and their interaction were performed with each CRF measure as the outcome variables for both V1 and V2/V3[1]. Consistent with our predictions, the model revealed a significant crossover interaction between categorization dimension and task phase in amplitude within area V1, $F$ (1, 23) = 9.87, $p$ = .003. Amplitudes were significantly higher during categorization compared to the contrast discrimination task among orientation rule learners, t (11) = 2.33, $p$ = .04, $d$ = 1.18. The spatial frequency group showed a trend in the opposite direction: amplitude on the contrast discrimination task was slightly greater and did not significantly differ from the categorization task, t (12) = -1.47, $p$ = .17, $d$ = 0.62 (Fig. 3). Between groups, orientation rule learners exhibited significantly higher amplitudes than spatial frequency rule learners during categorization, t (23) = 2.85, $p$ = .01, $d$ =1.25, but not during the orthogonal contrast discrimination task, t (23) = -1.33, $p$ = .20, $d$ = 0.53. Directionally consistent, albeit less reliable, patterns were observed in V2/V3 (categorization dimension × task phase interaction: $F$ (1, 23) = 3.52, $p$ = .07; categorization vs. contrast task: orientation rule learners: t (11) = 1.39, $p$ = .19, $d$ = 0.48; spatial frequency rule learners: t (12) = -.79, $p$ = .44, $d$ = 0.26; orientation vs. spatial frequency rule learners: categorization task: t (23) = 2.21, $p$ = .04, $d$ = 0.55; contrast task: t (23) = -.13, $p$ = .90, $d$ = .03).

Within V1, we similarly observed a significant two-way interaction between categorization dimension and task phase within slope, $F$ (1, 23) = 4.21, $p$ = .046. Slopes were significantly steeper in the categorization task compared to the contrast discrimination task for

[1] The results for areas V2 and V3 were closely matched across statistical comparisons, so the CRFs were averaged across both regions for all analyses.

the orientation group, t (11) = 2.78, $p$ = .02, $d$ =1.36, while they did not reliably differ for the spatial frequency group, t (12) = 1.08, $p$ = .30, $d$ = .50. Likewise, slopes were steeper for orientation rule learners compared to spatial frequency rule learners during categorization, t (23) = 2.94, $p$ = .007, $d$ = 1.27, but did not differ between the two groups during the contrast discrimination task, t (23) = .56, $p$ = .58, $d$ = 0.23. Once again, the slope patterns in V2/V3 were consistent but weaker than what was observed in V1 (categorization dimension × task phase interaction: $F$ (1, 23) = 3.23, $p$ = .08; categorization vs. contrast task: orientation rule learners: t (11) = 2.28, $p$ = .04, $d$ = 0.79; spatial frequency rule learners: t (12) = .66, $p$ = .52, $d$ = 0.18; orientation vs. spatial frequency rule learners: categorization task: t (23) = 2.72, $p$ = .01, $d$ = 0.68; contrast task: t (23) = .33, $p$ = .74, $d$ = .09).

In addition to amplitude and slope, we tested the effects of category learning on orientation CRF center shift and bandwidth. Center shift reflects the relative precision of orientation representations, where absolute values indicate how close the peak of the CRF is to the true orientation presented on a given trial. Bandwidth reflects the specificity of the representation in orientation space. We found a marginally significant two-way interaction between categorization dimension and task phase for center shift within V1, $F$ (1, 23) = 3.81, $p$ = .06. Pairwise comparisons revealed that the CRF centers were significantly closer to 0° during category learning than during the contrast discrimination task among orientation rule learners, t (11) = -3.25, $p$ = .008, $d$ = 0.98. This pattern, however, was absent among the spatial frequency rule learners, t (12) = -.82, $p$ = .43, $d$ = 0.23. CRF centers for the orientation group were also significantly closer to 0° when compared to the spatial frequency group during the categorization task, t (23) = -2.32, p = .03, $d$ = 0.62, while the groups did not differ during the contrast discrimination task, t (23) = .69, $p$ = .50, $d$ = 0.19. This suggests that orientation rule learners

exhibited more precise representations of presented orientations than spatial frequency rule learners specifically during category learning. This pattern was restricted to V1, however (V2/V3: interaction term: $F$ (1, 23) = .94, $p$ = .34). In contrast to the other measures, bandwidth was not significantly modulated by categorization condition or phase in either V1, $F$ (1, 23) = .60, $p$ = .45, or V2/V3, $F$ (1, 23) = .14, $p$ = .71.

Thus far, we have compared reconstructed representations of stimulus orientation across visually matched tasks. Taken together, the results suggest that neural representations of orientation were enhanced during active categorization, and only when orientation was the category-relevant dimension (Fig. 3). This was largely true in all tested areas (V1 and V2/V3), albeit stronger and more reliable in primary visual cortex.

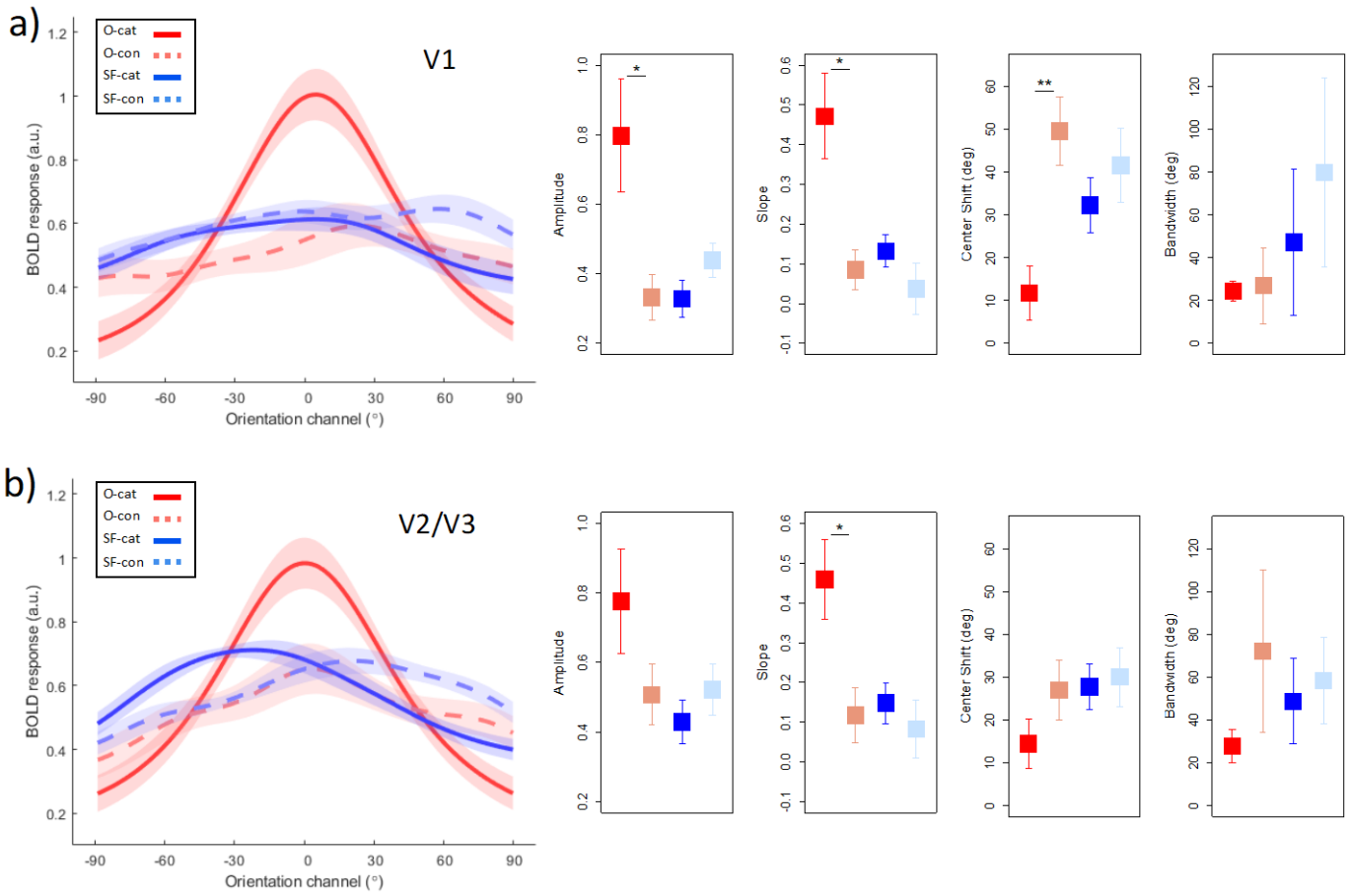


**Figure 3.** Effects of categorization and contrast discrimination on orientation CRFs. a) Mean orientation CRFs and parameter means for V1 and b) V2/V3. On the x-axes, 0° (center) corresponds to the true orientation value of the stimulus presented on a given trial. "O-cat" (red, solid lines) = orientation group, categorization task. "O-con" (faded red, dashed lines) = orientation group, contrast task. "SF-cat" (blue, solid lines) = spatial frequency group, categorization task. "SF-con" (faded blue, dashed lines) = spatial frequency group, contrast task. Error bands in the line plots reflect within-subject SEM. Error bars in the point plots reflect between-subject SEM. ** = $p < .01$, * = $p < .05$.

*Channel response functions for near-boundary orientations vs. central exemplars*

In the analyses thus far, we have considered all stimulus orientation values together. However, we hypothesized that representational enhancement should be most pronounced for difficult-to-classify stimuli that border the category boundary, where stronger and/or more specific perceptual representations would benefit performance the most. To test these

predictions, we created two groups of trials: ones containing orientation values near the assigned boundary (5° offset), and ones containing orientation values far from the boundary (35° and 45° offset).

Four linear mixed models with factors for categorization condition (orientation and spatial frequency), orientation offset from the category boundary (near and far), and their interaction were carried out for each tested visual area with amplitude, slope, center shift, and bandwidth as the respective outcome variables. The amplitude model revealed a significant interaction effect in V1 between category rule and stimulus distance from the boundary, $F$ (1, 23) = 5.56, $p$ = .03 (but not in V2/V3: $F$ (1, 23) = .62, $p$ = .44). Reconstructed representations of stimulus orientation had higher amplitudes for stimuli bordering the category boundary relative to those far from the boundary within the orientation group, t (11) = 4.06, $p$ = .002, $d$ = 0.76, but not the spatial frequency group, t (12) = 1.38, $p$ = .19, $d$ = 0.45.

The same interaction was significant within the slope of V1 CRFs, $F$ (1, 23) = 11.9, $p$ = .002: Within orientation rule learners, near-boundary stimuli elicited steeper slopes than those far from the boundary, t (11) = 4.32, $p$ = .001, $d$ = 0.88, an effect that was not present for spatial frequency rule learners, t (12) = -1.15, $p$ = .27, $d$ = 0.48 (Fig. 4). As with amplitude, this effect was largely restricted to V1 (V2/V3: $F$ (1, 23) = .71, $p$ = .41).

A converging albeit marginally significant two-way interaction between categorization task and distance from the boundary was present in center shift within V1, $F$ (1, 23) = 3.72, $p$ = .07 (but not in V2/V3: $F$ (1, 23) = .02, $p$ = .90). Consistent with predictions, CRFs were centered closer to the presented stimulus on near-boundary trials compared to far-from-boundary trials among orientation rule learners, t (11) = -2.60, $p$ = .02, $d$ = 0.45. At the same time, the centers did not significantly differ across distances from the arbitrary orientation boundaries among

561 spatial frequency rule learners, t (12) = .76, $p$ = .46, $d$ = 0.19. Furthermore, the mean center shift
562 was significantly closer to 0 on near-boundary trials among the orientation group compared to
563 the spatial frequency group, t (23) = -2.77, $p$ = .01, $d$ = 0.72. Consistent with our contrast
564 discrimination results (Fig. 3), categorization group and distance from the boundary did not
565 reliably modulate bandwidth in either V1, $F$ (1, 23) = .89, $p$ = .35, or V2/V3, $F$ (1, 23) = 2.70, $p$
566 = .11.

567 Taken together, our results across amplitude, slope, and center shift converge in strong
568 support of the hypothesis that sensory representations within V1 were made stronger and more
569 precise for task-relevant stimulus dimensions in response to learning. Moreover, this
570 enhancement was primarily applied to the most behaviorally relevant features in the space (in
571 this case, orientations flanking the category boundary).
572

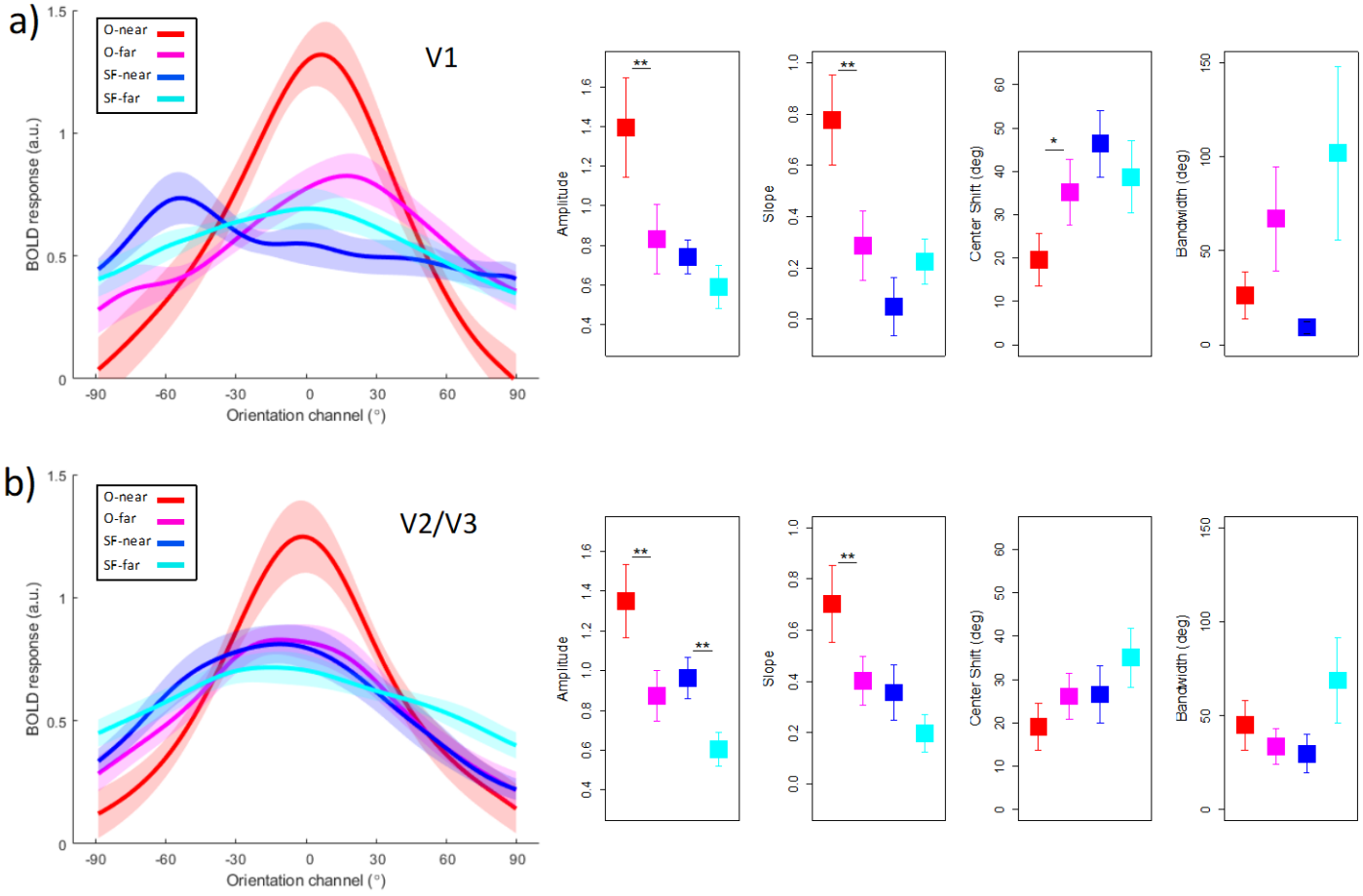


**Figure 4.** Categorization group and boundary effects on orientation CRFs. a) Mean orientation CRFs and parameter means for V1 and b) V2/V3. On the x-axes, 0° (center) corresponds to the true orientation value of the stimulus presented on a given trial. "O-near" (red) = orientation group, near-boundary stimuli. "O-far" (magenta) = orientation group, far-from-boundary stimuli. "SF-near" (blue) = spatial frequency group, near-boundary stimuli. "SF-far" (cyan) = spatial frequency group, far-from-boundary stimuli. Error bands in the line plots reflect within-subject SEM. Error bars in the point plots reflect between-subject SEM. ** = $p < .01$, * = $p < .05$.

*Effects of Learning on Reconstructed Representations of Orientation*

Over the course of 6 blocks of a categorization task, we demonstrated that category learning enhances the sensory representation of task-relevant features. One question that remains is whether and how these representations change over the course of learning. For example, it is possible that the boundary-specific enhancement of orientation representations in primary visual

588 cortex only emerges after asymptotic learning, when subjects have successfully detected and
589 established the location of the category boundaries. Alternatively, representational sharpening
590 may be driven by prediction error during active learning, and thus more apparent during the early
591 stages of the task when subjects may engage in explicit hypothesis testing to determine the
592 category rule (Choi et al., 1993; Johansen & Palmeri, 2002; Medin & Schaffer, 1978). Finally, a
593 third possibility is that the boundary-specific enhancement holds stable across the course of
594 learning. Most subjects across both categorization conditions reached a performance asymptote
595 by the fourth task block (see Fig. 2). Thus, to isolate possible early and late learning effects in
596 the present study, we compared reconstructed representations in blocks 1 and 2 of the
597 categorization task to those in blocks 5 and 6. Notably, the model training procedure was
598 identical for early- versus late-learning scanning runs.
599 To test whether the boundary-specific representational sharpening observed in V1 was
600 differentially modulated early or late during the learning process, we extended the previous
601 model to include a categorical predictor for early vs. late learning, with particular interest in the
602 3-way interaction between categorization condition, distance from the boundary, and learning
603 stage. In amplitude, this 3-way interaction was not significant, $F$ (1, 46) = .01, $p$ = .97. However,
604 the model revealed a significant 2-way interaction between categorization condition and
605 early/late learning, $F$ (1, 46) = 8.32, $p$ = .005. This interaction reflects the fact that across all
606 stimulus values, amplitudes were significantly higher in late versus early learning for the
607 orientation group, t (11) = 2.66, $p$ = .02, $d$ = 0.34, whereas learning duration was associated with
608 a significant decrease in amplitude among the spatial frequency group, t (12) = -2.20, $p$ = .048, $d$
609 = 0.63. This inverse effect in amplitude between the two groups occurred independently of
610 boundary effects: The difference in amplitude for near vs. far from boundary exemplars within

611 the orientation group did not differ between learning stages, t (11) = .01, $p$ = .99, $d$ = .001.
612 Neither the 3-way interaction effect, $F$ (1, 46) = .19, $p$ = .66, nor any marginal effects reached
613 significance when slope was used as the outcome variable.

614 Interestingly, orientation CRF bandwidths were also modulated in response to category
615 learning, reflected in a two-way interaction between categorization condition and learning stage,
616 $F$ (1, 46) = 5.80, $p$ = .02. Independent of category boundaries, bandwidths were relatively
617 narrower in late learning compared to early learning in the orientation group, t (11) = -1.91, $p$ =
618 .08, $d$ = 1.02, albeit not significantly so. At the same time, we observed a learning effect trending
619 in the opposite direction for the spatial frequency group, t (12) = 2.01, $p$ = .07, $d$ = 0.69, whereby
620 bandwidths widened over the course of learning. Across the board, patterns in V2/V3 were once
621 again directionally consistent with V1, but largely unreliable (see Fig. 5).

622 The combined amplitude and bandwidth modulation observed for V1 suggests that
623 representations of category-relevant stimulus dimensions are enhanced, especially at later stages
624 of learning. At the same time, boundary-specific representational changes emerged relatively
625 early in learning and remained consistent after subjects reached asymptotic performance.
626
627

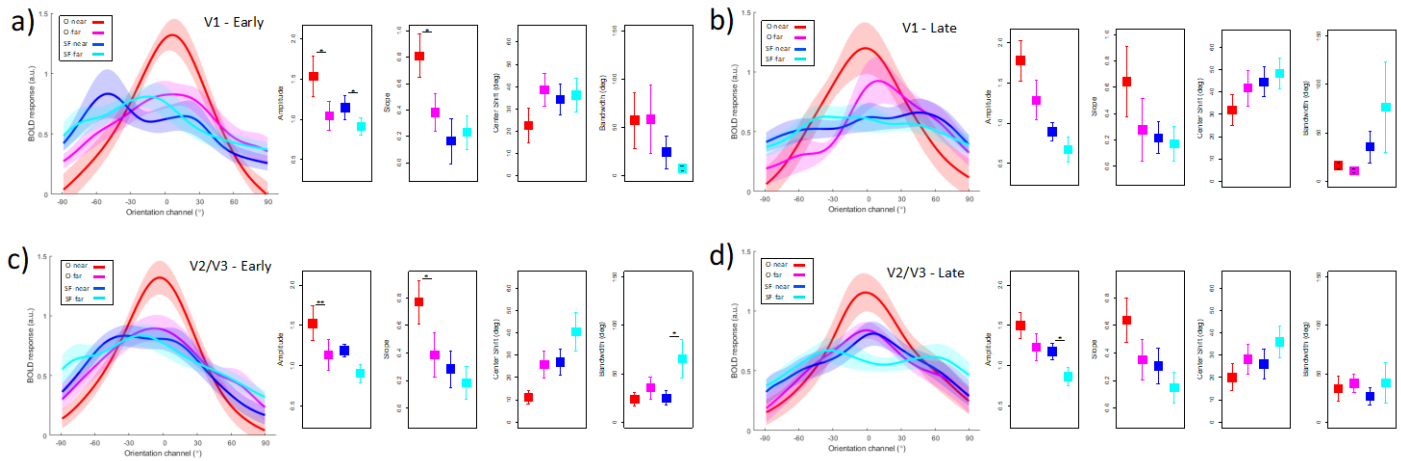


**Fig. 5.** Categorization group and boundary effects on orientation CRFs divided into early and late learning stages. a) Mean orientation CRFs and parameter means for V1 during early learning, b) V1 during late learning, c) V2/V3 during early learning, and d) V2/V3 during late learning. Error bands in the line plots reflect within-subject SEM. Error bars in the point plots reflect between-subject SEM. ** = $p < .01$, * = $p < .05$.

*Association Between Task Accuracy and Reconstructed Representations of Orientation*

Finally, we were interested in testing whether differences in behavioral performance between subjects during the scanning session were associated with the shape of their reconstructed representations of orientation specifically for boundary-adjacent exemplars. On one hand, it is possible that representational sharpening was most pronounced for high-performing subjects who had more time to narrow their attentional focus on the category boundaries after quickly establishing the category rule. Alternatively, it is possible that the representational sharpening observed for near-boundary exemplars in the orientation group is an error-driven effect, such that individuals who were committing more errors in this perceptually challenging region of the feature space would exhibit the strongest sharpening effects as a compensatory mechanism.

To address this question, we performed Pearson correlations between mean categorization accuracy and subject-level differences in CRF parameters associated with near- versus far-from-boundary stimuli (e.g., near amplitude − far amplitude). Because all observed patterns were stronger and more reliable in primary visual cortex, we expected any significant associations with behavior to occur in this area. We found that orientation subjects' categorization accuracy was significantly associated with relatively higher CRF amplitudes in V1 for near-boundary stimuli, $r = .59$, t (10) = 2.31, $p = .04$, in addition to relatively steeper slopes for near-boundary stimuli, $r = .68$, t (10) = 2.90, $p = .02$. Among spatial frequency learners, the associations between task accuracy and indices of near-boundary representational enhancement were negative and non-significant (amplitude: $r = -.46$, t (11) = -1.71, $p = .12$; slope: $r = -.10$, t (11) = -.33, $p = .75$). We found no significant associations between learning performance in V1 center shift or bandwidth, nor among any individual CRF parameters in V2/V3.

Collectively, the results suggest that learning category rules defined by orientation not only led to sharper representations of orientation than learning an orthogonal rule, but that the strength and specificity of the reconstructed representations in V1 for challenging near-boundary (5°) exemplars track individual differences in categorization accuracy. Higher-performing orientation subjects exhibited more relative enhancement of near-boundary orientation representations than lower-performing subjects who nonetheless learned the category rules.

## Discussion

Learning to categorize visual stimuli leads to improved perceptual representations (Hamm et al., 1998; Jolicoeur et al., 1984; Zeithamova & Maddox, 2007; Soto & Ashby, 2015),

669 but a clear consensus has not been reached on how and when this occurs. We hypothesized that
670 category learning is supported by a representational enhancement of near-boundary exemplars
671 within sensory cortex, and that this effect should be most pronounced *during* learning, especially
672 if it serves to facilitate exploration and discovery of the boundaries that define one category from
673 another. Using task manipulations that matched all aspects of the visual display, we found
674 evidence of stronger and sharper orientation representations during a categorization task
675 compared to an orthogonal contrast discrimination task and only for orientation rule learners.
676 Moreover, reconstructed representations of near-boundary exemplars within V1 exhibited higher
677 amplitudes, steeper slopes, and smaller shifts from the presented orientation compared to those
678 far from the boundary. This suggests that visual category learning is accompanied by rapid,
679 feature-specific functional plasticity in early visual cortex to support more challenging category-
680 relevant perceptual discriminations.

681 Whether neural plasticity generalizes to early visual cortex during categorization has been
682 largely ignored or discounted in most neuroscientific investigations (Freedman & Assad, 2016).
683 Although the importance of accounting for learning-related attentional flexibility has long been
684 recognized across theoretical accounts of categorization (Nosofsky, 1986; Kruschke, 1992), most
685 neuroimaging research has focused on higher-order visual areas (Li et al., 2007; Meyers et al.,
686 2008; Folstein et al., 2013; Mack et al., 2013; Uyar et al., 2016; O'Bryan et al., 2018). Collectively,
687 these demonstrations show that extrastriate occipital, temporal, parietal, and frontal regions exhibit
688 greater sensitivity to stimuli along attended, diagnostic dimensions relative to those that do not
689 predict category membership. Similarly, patterns in ventral occipitotemporal cortex contain abstract
690 representations that distinguish between learned categories over and above sensory properties alone
691 (Li et al., 2007; Meyers et al., 2008).

Mounting evidence indicates that attentional control can exert modulatory effects in early visual cortex, including V1 (Kamitani & Tong, 2005; Scolari et al., 2012; Scolari & Serences, 2009; Serences et al., 2009), leaving open the possibility that V1 may contribute to category learning despite its limited treatment in the literature. Moreover, perceptual learning elicits sensory modulation in early visual cortex by increasing gain for attended features, suggesting that such learning may directly support behaviorally relevant perceptual discriminability (Byers & Serences, 2014).

In line with these findings, the current study provides another demonstration that neural representations in early, sensory-driven regions can be rapidly and robustly modified to optimize behavior in a learning context. Our results compliment recent work by Ester and colleagues (2020), who used an inverted encoding model to test for categorization-related sensory modulation in visual cortex. Participants in their study were trained to ceiling performance on an orientation categorization task prior to scanning, such that their fMRI results reflect perceptual representations following, but not during the acquisition of category rules. The results revealed that the reconstructed representations of stimulus orientation were shifted towards the mean of the category they correctly belonged to after learning was complete, suggesting that increasing within-category similarity in sensory populations supports generalization of learned categorization.

This study extends the previous findings by providing novel support for localized representational enhancement during ongoing category learning. Stimuli bordering orientation rule learners' assigned boundaries elicited stronger (via increased amplitude), more specific (via steeper slope), and more faithful (via center shift) representations of orientation values in visual cortex than far-from-boundary stimuli. This was especially true in V1. This representational gain implies a stretching of the feature space that is specific to difficult-to-classify exemplars. For

example, small offset differences near an orientation category boundary should be more neurobiologically separable (and by extension, perceptually separable) than an identical difference between exemplars near a category center. Moreover, among orientation rule learners, the reconstructed representations for stimulus orientation falling at least 35 degrees from a boundary did not significantly differ from those observed in two orthogonal tasks (contrast discrimination; spatial frequency categorization) where stimulus orientation was irrelevant to task performance. These results suggest that representational sharpening – at least for orientation perception – can be a highly specific effect.

Categorization behavior is regularly modeled with great success, especially via exemplar models, which posit that observers classify new stimuli based on their similarity to memory representations of previously encountered category exemplars (Rodrigues & Murre, 2007). The most widely used exemplar-based computational models apply attentional weights uniformly across feature values depending on dimensional relevancy (Goldstone & Steyvers, 2001; Op de Beeck et al., 2003; Gureckis & Goldstone, 2008; Nosofsky, 2011; Ashby & Rosedahl, 2017). Strong, positive attentional weighting then leads to dimension-wide perceptual stretching-- an outcome that has received support among behavioral studies that fail to detect localized perceptual effects (Folstein et al., 2012; Folstein et al., 2014; Van Gulick & Gauthier, 2014). Nonetheless, few studies have demonstrated that more flexible models which allow for exemplar-specific attentional modulation may do a better job of accounting for human behavior in certain categorization tasks (Aha & Goldstone, 1992; Rodrigues & Murre, 2007). Our data provide evidence that, in the context of a categorization task, flexible modulation is possible. However, in a departure from traditional models, we demonstrate this may not operate solely on memory representations, per se, but

involves low-level sensory areas. This might be especially true for stimuli that can be classified based on feature values within a single, continuous dimension.

Participants may initially apply top-down attention during early category learning in service of explicit hypothesis testing, especially in the context of simple (unidimensional) rules (Choi et al., 1993; Johansen & Palmeri, 2002; Medin & Schaffer, 1978). This is consistent with our interpretation of the current data. Johansen and Palmeri further demonstrated that as learning progresses, participants tend to switch from a rule-based approach to an exemplar-based one. Because we assessed orientation representation during learning, it is possible the boundary-specific enhancement we observed would dissipate after participants settled on a precise categorization rule. This may then give way to a representational shift towards category center exemplars (Ester et al., 2020) to maximize accurate classification of novel members. Although we did not observe this same center shift in the current study, our results did reveal a relative increase in amplitude for central exemplars during late learning stages, which could potentially serve as a precursor to an exemplar-based approach.

Relatedly, more research is needed to establish how categorization-related perceptual enhancements transfer to orthogonal tasks and novel stimuli after subjects are no longer actively engaged in category learning. Different studies attempting to characterize such sensory modulations at the neural and behavioral levels have tested their predictions during the learning process (Goldstone & Steyvers, 2001; Sigala & Logothetis 2002), following asymptotic learning (Ester et al., 2020; Folstein et al., 2012; 2013; Jiang et al., 2007), and using interleaved categorization and discrimination tasks (Gureckis & Goldstone, 2008). Future studies should seek to contrast these different approaches to establish which scenarios best facilitate transfer

between category learning and discrimination performance, or neural indices of perceptual sensitivity.

Behavioral perceptual effects associated with category learning may reflect either plasticity in sensory populations tuned to relevant stimulus features (sensory modulation; Sigala & Logothetis, 2002; Treue & Martinez-Trujillo, 2004; Yang & Manusell, 2004), a more efficient transmission of visual information to higher-level regions implicated in executing decisions based on sensory input (enhanced readout; Dosher & Lu, 1999, 2009; Lu & Dosher, 1999; Law & Gold, 2008; Freedman & Assad, 2016), or both. Here, we have provided evidence of task-specific modulations within early visual cortex in support of category learning, and this may be supported by attentional control guidance from frontoparietal cortex and/or signals from frontal regions that are routinely activated during categorization tasks (e.g., rostrolateral prefrontal cortex; Davis, et al., 2017; O'Bryan et al., 2018b; Paniukov & Davis, 2018). Future research should further explore whether and how higher-order regions interact with sensory cortex in support of category learning.

In conclusion, our data support the prediction that visual category learning is associated with a representational sharpening in sensory populations that are tuned to category-relevant stimulus dimensions. We additionally showed that such sharpening was uniquely observed for challenging stimuli that bordered subjects' category boundaries. Collectively, these results suggest that learning-related changes to the human visual system may be implemented more flexibly and efficiently than previously thought.

782

783

784 **References**

785 Arcaro, M.J., McMains, S.A., Singer, B.D., & Kastner, S. (2009). Retinotopic organization of
786 human ventral visual cortex. *J Neurosci,* 29: 10638 – 10652.
787 Brouwer, G., & Heeger, D. (2009). Decoding and reconstructing color from responses in human
788 visual cortex. *J Neurosci* 29: 13992–14003.
789 Brouwer, G., & Heeger, D. (2011). Cross-orientation suppression in human visual cortex. *J*
790 *Neurophysiol* 106: 2108–2119.
791 Byers, A. & Serences, J.T. (2014). Enhanced attentional gain as a mechanism for generalized
792 perceptual learning in human visual cortex. *J. Neurophys. 112*, 1217-1227.
793 Davis, T., Goldwater, M.B., & Giron, J. (2017). From concrete examples to abstract relations:
794 The rostrolateral prefrontal cortex integrates novel examples into relational categories.
795 *Cereb Cortex, 27*, 2652-2670.
796 Dosher, B.A., & Lu, Z.L. (1999). Mechanisms of perceptual learning. *Vision Research* 39: 3197–
797 3221.
798 Dosher, B.A., & Lu, Z.L. (2009). Hebbian reweighting on stable representations in perceptual
799 learning. *Learn Percept* 1: 37–58.
800 Ester, E.F., Sprague, T.C., & Serences, J.T. (2020). Categorical biases in human
801 occipitotemporal cortex. *The Journal of Neurocience*. 40: 917 – 931.
802 Folstein, J., Newton, A. Van Gulick, A.B., Palmeri, T., & Gauthier, I. (2012). Category learning
803 causes long-term changes to similarity gradients in the ventral stream: A multivoxel
804 pattern analysis at 7T. *Journal of Vision 12 (9),* 1106-1106.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cereb Cortex*, *23*, 814–823.

Folstein, J.R., Palmeri, T.J., Van Gulick, A.E., & Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Current directions in Psychological Science, 24*, 17-23.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*, 5235–5246.

Freedman, D.J., & Assad, J.A. (2016). Neuronal mechanisms of visual categorization: an abstract view on decision making. *Annual Review of Neuroscience*, *39*, 129-147.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*, 116–139.

Jiang, X., Bradley, E., Rini, R.A., Zeffiro, T., VanMeter, J., & Riesenhuber, M., (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron, 53*, 891–903.

Kamitani, Y., & Tong, F. (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci, 8*, 679 – 685.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psych Rev*, *99*, 22-44.

Law, C.T., & Gold, J.I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nat Neurosci* 11: 505–513.

Li, S., Ostwald, D., Giese, M., & Kourtzi, Z. (2007). Flexible coding for categorical decisions in the human brain. *J. Neurosci., 27*, 12321-12330.

828 Ling, S., Liu, T., Carrascim M. (2009). How spatial and feature-based attention affect the gain
829 and tuning of population responses. *Vision Research, 49*(10), 1194-1204.
830 Lu, Z.L., & Dosher, B. (1999). Characterizing human perceptual inefficiencies with equivalent
831 internal noise. *J Opt Soc Am A Opt Image Sci Vis* 16: 764–778.
832 Mack, M.L., Preston, A.R., & Love, B.C. (2013). Decoding the brain's algorithm for
833 categorization from its neural implementation. *Current Biology*, *23*, 1-5.
834 Martinez-Trujillo, J. C. and Treue, S. (2004). Feature-based attention increases the selectivity of
835 population responses in primate visual cortex. *Current Biology, 14*, 744–751
836 Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., & Poggio, T. (2008). Dynamic
837 population coding of category information in inferior temporal and prefrontal cortex. *J.*
838 *Neurophysiol., 100*, 1407-1419.
839 Nosofsky, R.M. (1992). Similarity scaling and cognitive process models. *Annu. Rev. Psychol.* 43,
840 25–53.
841 O'Bryan, S.R., Walden, E., Serra, M.J., & Davis, T. (2018). Rule activation and ventromedial
842 prefrontal engagement support accurate stopping in self-paced learning. *NeuroImage*,
843 *172*, 415-426.
844 O'Bryan, S.R., Worthy, D.A., Livesey, E.J., & Davis, T (2018). Model-based fMRI reveals
845 dissimilarity processes underlying base rate neglect. *eLife*, *7*, 46395.
846 Op de Beeck, H. P., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the
847 representation of shape: Dimensions can be biased but not differentiated. *Journal of*
848 *Experimental Psychology: General*, *132*, 491–511.
849 Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*,
850 *5*, 291-303.

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

Scolari, M. & Serences, J. T. (2009). Adaptive allocation of attentional gain. *Journal of Neuroscience, 29*, 11933–11942.

Scolari, M., Byers, A., & Serences, J. T. (2012). Optimal deployment of attentional gain during fine discriminations. *Journal of Neuroscience, 32*, 7723–7733.

Scolari, M., Ester, E. F. & Serences, J. T. (2014). Feature- and object-based attentional modulation in the human visual system. In: *The Oxford Handbook of Attention* (A.C. Nobre & S. Kastner, Eds).

Serences, J.T., Saproo, S., Scolari, M., Ho, T., & Muftuler, L.T. (2009). Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *Neuroimage, 44*, 223–231.

Serences JT, & Saproo S. (2012) Computational advances towards linking BOLD and behavior. *Neuropsychologia, 50*, 435–446.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*, 318–320.

Soto, F.A., & Ashby, F.G. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition, 139*, 105-129.

Sprague, T.C., & Serences, J.T. (2015). Using human neuroimaging to examine top-down modulation of visual perception. In: Forstmann, B., Wagenmakers, E.J. (eds) An introduction to model-based cognitive neuroscience. Springer, New York, NY.

Sprague, T.C., Adam, K.C.S., Foster, J.J, Rhamati, M., & Serences, J.T. (2018). Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. *Eneuro, 5*, 3.

Swisher, J.D., Halko, M.A., Merabet. L.B, McMains, S.A., & Somers, D.C. (2007).Visual topography of the human intraparietal sulcus. *The Journal of Neuroscience, 20*, 5326 – 5337.

Treue, S. and Martinez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature 399*, 575–579.

Uyar, F., Shomstein, S., Greenberg, A.S., & Behrmann, M. (2016). Retinotopic information interacts with category selectivity in human ventral cortex. *Neuropsychologia*. http://dx.doi.org/10.1016/j.neuropsychologia.2016.05.022i

Yang, T., & Maunsell, J. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *J Neurosci* 24, 1617–1626.