Occipitotemporal Representations Reflect Individual Differences in Conceptual Knowledge

Braunlich, K.[a,*], Love, B. C.[a,b]

[a] *Department of Psychology, University College London. 26 Bedford Way, London, UK.*
[b] *The Alan Turing Institute. 96 Euston Rd., London, UK.*

Abstract

Through selective attention, decision-makers can learn to ignore behaviorally-irrelevant stimulus dimensions. This can improve learning and can increase the perceptual discriminability of relevant stimulus information. Across cognitive models of categorization, this is typically accomplished through the inclusion of attentional parameters, which provide information about the importance of each stimulus dimension during decision-making. These parameters are often described geometrically, such that perceptual differences over relevant psychological dimensions are accentuated (or stretched), and differences over irrelevant dimensions are down-weighted (or compressed). An open question is whether these attention parameters are reflected in the brain. We examined two published fMRI datasets in which participants learned to categorize stimuli with features that differed in their relevancy. Attentional parameters were derived for each participants using formal categorization models fit to behavioral data. Using multivoxel pattern analysis (MVPA), we found that when greater attentional weight was devoted to a particular visual feature (e.g., "color"), its value (e.g., "red") was more easily decoded from occipitotemporal cortex. This effect was sufficiently sensitive to reflect individual differences in conceptual knowledge, indicating that occipitotemporal stimulus representations are embedded within a space similar to that formalized by classic categorization theory.

Keywords: Concepts, Selective Attention, Occipitotemporal Cortex

Conflict of interest: The authors declare no conflicts of interest.

Acknowledgments: We thank the authors of the original studies for sharing their data. This work was supported by NIH Grant 1P01HD080679, Leverhulme Trust grant RPG-2014-075 and Wellcome Trust Senior Investigator Award WT106931MA to BCL.

*Corresponding author
Email address: k.braunlich@ucl.ac.uk (Braunlich, K.)

Introduction

Through selective attention, knowledge of abstract concepts can emphasize relevant stimulus features. For example, while the size of garments is critical when choosing what to purchase, weight may be more important when deciding how to ship them. The attention devoted to individual features is flexibly-modulated according to current goals, transient contextual demands, and is known to reflect evolving conceptual knowledge (Tversky, 1977; Goldstone, 2003). In formal categorization models, a way to account for this flexibility is through inclusion of attentional parameters, which reflect the influence of each dimension on the category decision (e.g., Kruschke, 1992; Nosofsky, 1986; Love et al., 2004). These attentional parameters are often described as "warping" psychological space, such that differences along relevant stimulus dimensions are accentuated (or "stretched") and differences along irrelevant dimensions are down-weighted (or "compressed"; Figure 1). Here, we directly test this classic idea by investigating whether the strength of neural representations of stimulus features are modulated by these attentional parameters.

When identifying specific objects, agents must typically consider all stimulus features, and the psychological distance between stimuli closely reflects their perceptual attributes (Shepard, 1957; Townsend & Ashby, 1982). During categorization, however, groups of stimuli must be treated equivalently, and both learning and generalization can be improved by selectively attending to relevant stimulus dimensions (Shepard et al., 1961; Nosofsky, 1986). Although categorization models differ in how stimuli are represented in memory (e.g., as individual exemplars, Nosofsky 1987; as prototypes Nosofsky 1987; Minda & Smith 2002; Zaki et al. 2003; or as clusters that flexibly mirror environmental structure, Anderson 1991; Love et al. 2004) they similarly assume that category learning involves learning to distribute attention across stimulus features so as to optimize behavioral performance. Although they differ in their mathematical details, these models also posit that *endogenous* ("top-down",or goal-directed) attentional control (Miller & Cohen, 2001; Tsotsos, 2011) attentional factors modulate the influence of *exogenous* (or perceptual) stimulus dimensions on behavioral choice. When fit to behavioral data, the attentional parameters, therefore, provide interpretable information about the strategies used by individual decision-makers to integrate information from the external world.

In the brain, endogenous attentional control is thought to involve modulation of sensory and association cortex by frontal regions representing abstract factors, such as goals, and task-set (Miller & Cohen, 2001; Buschman et al., 2012; Wallis et al., 2001; Baldauf & Desimone, 2014; Bichot et al., 2015; Gregoriou et al., 2014; Schwedhelm et al., 2017). The ventral prearcuate sulcus, inferior frontal gyrus, and frontal-eye fields may play a particularly important role in the feature-based attentional modulation of occipitotemporal cortex (Baldauf & Desimone, 2014; Bichot et al., 2015; Gregoriou et al., 2014). Effects associated with endogenous

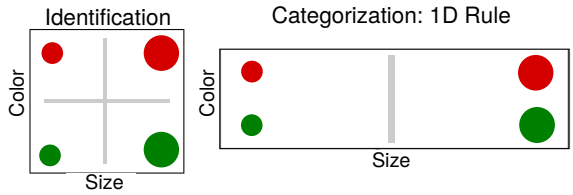

Figure 1: **Example: Attention Influences Psychological Space**. **Left:** In this object identification task, psychological dimensions should receive equivalent attention, as all stimulus dimensions are equally relevant. **Right:** In this example of rule-based categorization, only a single dimension is relevant (size), and decision-makers could ignore the irrelevant dimension (color). This is often described as "warping" psychological space such that differences along relevant dimensions are accentuated (or "stretched"), and differences along irrelevant dimensions are down-weighted (or "compressed").

attention are apparent across the visual cortical hierarchy (Buffalo et al., 2010; Luck et al., 1997; Motter, 1993; Poort et al., 2015); and several groups have noted that, with categorization training, representations of relevant stimulus dimensions in occipitotemporal cortex tend to be more discriminable than those that are behaviorally-irrelevant (Li et al., 2007; Folstein et al., 2013; Dieciuc et al., 2017; Sigala & Logothetis, 2002). This implies that occipitotemporal representational space may closely resemble that conceptualized by classic cognitive theory. Specifically, its expansion and and contraction, along axes defined by perceptually-separable dimensions (Garner, 1976), might closely reflect the cognitive strategies used by individual participants during decision-making.

To investigate this possibility, we combined model-based fMRI (Turner et al., 2017; Palmeri et al., 2017) with multivoxel pattern analyses (MVPA). This allowed us to link brain and behavior via the latent attentional parameters of the cognitive models. We applied this analysis to two publicly available datasets (https://osf.io). In the first (Mack et al., 2013), participants categorized abstract stimuli that varied according to four binary dimensions (Figure 2.A), according to a categorization strategy they learned prior to scanning. In the original paper, the authors fit both the Generalized Context Model (GCM; Nosofsky, 1986) and the Multiplicative Prototype Model (MPM; Nosofsky, 1987; Nosofsky & Zaki, 2002) to the behavioral data, and used them to compare exemplar and prototype accounts of occipitotemporal representation. In the second dataset (Mack et al., 2016), participants learned, while scanning, to categorized images of insects that varied according to three binary perceptual dimensions (Figure 2.B). The authors fit the SUSTAIN learning model (Supervised and Unsupervised STratified Adaptive Incremental Network; Love et al., 2004) to the behavioral data, and used it to investigate hippocampal involvement in the development of new conceptual knowledge.

As each dataset employed stimuli that varied according to multiple binary perceptual dimensions, we were able to investigate how these dimensions were concurrently represented in occipital cortex. In order to compare our results to those of the original papers, we used the GCM for the the first dataset (i.e., the winning model

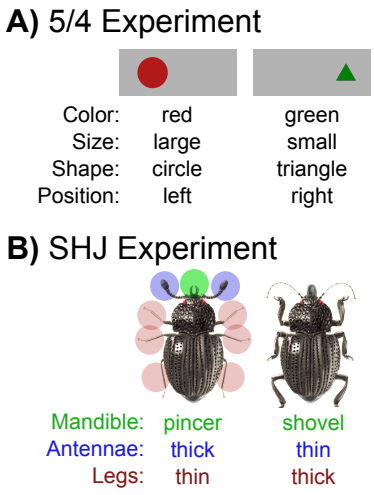

Figure 2: **Stimuli. A)** Two of the 16 stimuli used in the 5/4 experiment are illustrated. The stimuli varied according to four binary perceptual dimensions: color, size, shape and position. **B)** Two of the eight stimuli used in the SHJ experiment are illustrated. The stimuli were pictures of insects that varied according to three binary dimensions: mandible shape (highlighted in green), antennae thickness (highlighted in blue), and leg thickness (highlighted in red). For both experiments, the mapping of visual dimension to its role in each category structure (Tables 1 and 2) was randomized for each participant.

from Mack et al. 2013). In the second dataset, we considered the attentional parameters derived from SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network; Love et al., 2004). Although these models differ in how stimuli are represented (i.e., as exemplars, prototypes, or clusters), and differ in the way attentional weights are calculated, they similarly posit that attention "warps" psychological space, as illustrated in Figure 1. Thus, we aim to directly evaluate whether attention weights in successful cognitive models reflect neural representations both at the group and individual level.

Methods

Description of Datasets

In both experiments, interactions between category structure and individual differences in conceptual knowledge determined the relevance of each visual attribute to the categorization decision. The mapping of perceptual attributes to their role in the category structure was randomized for each participant. This allowed differentiation of effects associated with intrinsic perceptual stimulus properties from those associated with latent conceptual properties. For example, color may be relevant to categorization for one participant and irrelevant for another, allowing us to assess how neural coding differs as a function of endogenous attention apart from exogenous stimulus factors. In both experiments, participants learned to categorize the stimuli into one of two novel categories through trial-and-error.

The 5/4 Dataset

The first dataset (Mack et al., 2013) was collected while 20 participants (14 Female) categorized abstract stimuli (Figure 2.A), which varied according to four binary stimulus dimensions (size: large vs. small, shape: circle vs. triangle, color: red vs. green, and position: left vs. right). Prior to scanning, they learned, through trial-and-error to categorize the stimuli according to the 5/4 categorization task (Medin & Schaffer, 1978). During training, participants observed only the first nine stimuli shown in Table 1 (i.e., five category "A" members: A1-A5 and four category "B" members: B1-B4), and experienced 24 repetitions of each stimulus. During scanning, they were required to additionally categorize the seven transfer stimuli (i.e., T1-T7). In the scanner, stimuli were presented for 3.5 seconds on each trial, no feedback was provided, and stimuli were separated by a 6.5 second intertrial interval. Over six runs, each of the 16 stimuli were presented three times. The order of the stimulus presentations were randomized for each participant.

Stimulus	d1	d2	d3	d4
A1	1	0	0	0
A2	1	0	1	0
A3	0	1	0	0
A4	0	0	1	0
A5	0	0	0	1
B1	1	1	0	0
B2	1	0	0	1
B3	0	1	1	1
B4	1	1	1	1
T1	0	1	1	0
T2	1	1	1	0
T3	0	0	0	0
T4	1	1	0	1
T5	0	1	0	1
T6	0	0	1	1
T7	1	0	1	1

Table 1: **The 5/4 Category Structure**. Prior to scanning, participants learned to categorize the binary stimuli illustrated in Figure 2.A, via trial and error, using the first nine stimuli (category "A": A1-A5; category "B": B1-B4). During scanning, participants categorized both the training and the transfer (T1-T7) stimuli. Perceptual stimulus dimensions (see Figure 2) were randomly assigned for each participant to this abstract category structure.

The SHJ Dataset

In the second dataset (Mack et al., 2016), 23 right-handed participants (11 Female, mean age = 22.3 years) categorized images of insects (Figure 2.B) varying along three binary dimensions (legs: thick vs. thin, antennae: thick vs. thin, and mandible: pincer vs. shovel). We excluded data from two participants who each had corrupted data on one run. This resulted in 21 participants for the final analyses. During scanning, participants learned to categorize the stimuli according to the type one, type two, and type six problems described by Shepard et al.

(1961). In the type one problem, the optimal strategy required attending to a single stimulus dimension (e.g., "legs") that perfectly predicted the category label, while ignoring the other two dimensions. In the type two problem, the optimal strategy was a logical XOR rule, in which two stimulus features had to be considered together. In the type six problem, all stimulus features were relevant to the decision, and participants had to learn the mapping between individual stimuli and the category label. To maximally differentiate endogenous and exogenous factors, the irrelevant feature in the type two rule was used as a relevant feature of the type one problem for each participant.

Each problem was performed for four scanner runs. While all participants learned to perform the type six problem first, the order of the type one and type two problems was then counterbalanced across participants. Each trial consisted of a 3.5 second stimulus presentation period, a jittered 0.5-4.5 second fixation period. Feedback was presented for 2 seconds and consisted of an image of the presented insect and text indicating whether the response was correct or incorrect. Each trial was separated by jittered intertrial interval (4-8 seconds) consisting of a fixation cross.

For consistency across datasets, we used the group-derived ROI used in 5/4 dataset (Figure 3.B), and performed a similar analysis. As participants in the SHJ experiment learned to perform the type one, type two and type six problems during scanning, we mirrored the strategy used by the original authors, and divided the scanning sessions into early (first two runs of each problem) and late learning epochs (last two runs of each problem). We investigated the relationship between occipitotemporal representation and attention only during this late learning phase, where behavior had largely stabilized.

Stimulus	d1	d2	d3	Type1	Type2	Type6
1	0	0	0	A	A	A
2	0	0	1	A	B	B
3	0	1	0	A	B	B
4	0	1	1	A	A	A
5	1	0	0	B	A	B
6	1	0	1	B	B	A
7	1	1	0	B	B	A
8	1	1	1	B	A	B

Table 2: **SHJ Category Structures.** Participants learned by trial-and-error to perform the type one (a one-dimensional rule-based categorization task), type two (a two-dimensional XOR rule-based categorization task), and type six (a three-dimensional task requiring memorization of the individual stimuli) problems during scanning. Perceptual stimulus dimensions (see Figure 2) were randomly assigned for each participant to this abstract category structure.

Image Processing

Preprocessing included motion correction, and coregistration of the anatomical images to the mean of the functional images (using SPM, version 6470). All MVPA analyses were performed in native space without smoothing. For group-level analyses, the statistical maps from each participant were warped to MNI space using ANTs (Avants et al., 2009), and then smoothed with a 6mm full-width at half maximum Gaussian kernel. The ROI derived from group-level analyses were transformed back into native space for ROI-level analyses. We performed MVPA on the unsmoothed, single-trial, *t*-statistic images (Misaki et al., 2010) derived from the least-squares separate procedure (LSS; Mumford et al., 2012). We used SPM to estimate the LSS images for the 5/4 dataset, but used the NiPy python package (http://nipy.org/nipy/index.html) for the SHJ dataset, as it tends to run more efficiently, and this study used a multiband sequence with smaller voxel dimensions.

Results

Representations of Individual Visual Features

We chose to focus on regions most strongly representing the exogenous stimulus features. Representations in visual cortex are better understood than those in higher order cortex, and their representations of exogenous stimulus features are known to be modulated by selective attention (similarly to the effect illustrated in Figure 1 Li et al., 2007; Folstein et al., 2013; Dieciuc et al., 2017; Sigala & Logothetis, 2002; Sigala, 2004). To identify these regions, we performed a cross-validated searchlight analysis (sphere radius = 10mm.; Kriegeskorte et al., 2006)[1] in which we decoded each of the four visual features (position, shape, color and size). We performed the analysis in native anatomical space, using a linear support vector classifier (SVC, C=0.1; using the Scikit-learn python package; Pedregosa et al., 2011) in conjunction with a five-fold, leave-one-run out, cross-validated procedure. This involved repeatedly training the model on four of the five runs, and testing whether it could accurately predict the stimulus features associated with the held-out neuroimaging data.

After centering each of the resultant statistical maps at chance (50% for each visual feature), we created a single map for each participant, reflecting the average decoding accuracy across features. We then normalized each map to MNI space, and performed a group-level permutation test. This involved randomly flipping the sign of the statistical maps 10,000 times (using FSL's randomise function; Winkler et al., 2014). The familywise error rate was controlled using a voxelwise threshold of $p < 0.001$. This identified right middle frontal gyrus (BA9) and left postcentral motor cortex, as well as widespread visual and association cortex, extending dorsally from

[1]This involves moving an imaginary sphere throughout the brain; repeatedly investigating how well the voxels could decode of the features presented on each trial.

occipital pole to the bilateral superior extrastriate cortex and bilateral intraparietal sulcus (IPS), and ventrally into the bilateral lingual gyrus (Table 3). As this procedure yielded a diffuse pattern of spatial activity, we increased the minimum *t*-statistic threshold (from 6.24 to 9) to isolate voxels most strongly representing the exogenous stimulus dimensions. This removed voxels belonging to the bilateral inferior occipital cortex (LOC), left lingual gyrus, bilateral IPS, and bilateral precuneus, and thus excluded association and higher-order sensory cortices. The resultant ROI is illustrated in Figure 3.A (yellow).

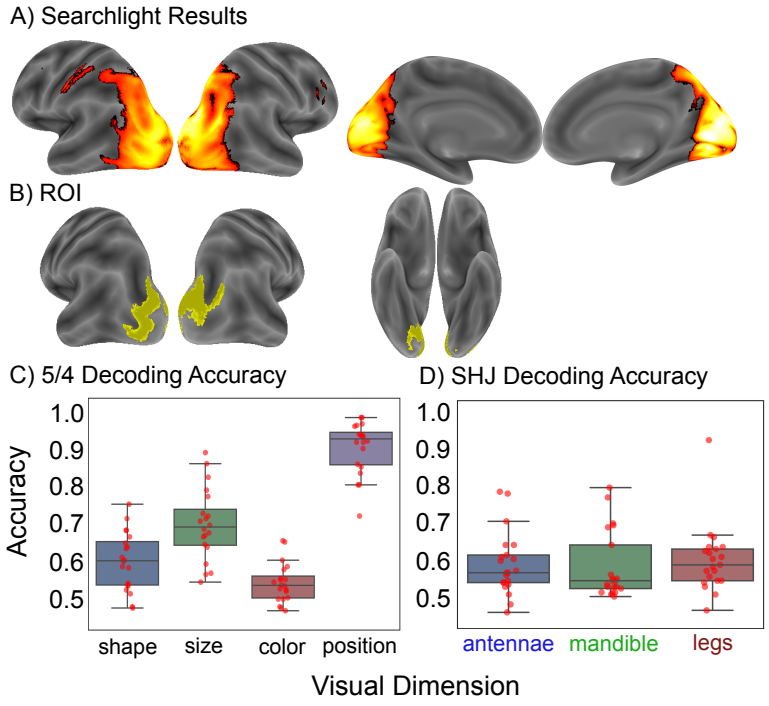

Figure 3: **A)** For the 5/4 dataset, a searchlight analysis indicated that binary perceptual dimensions could be decoded from widespread visual regions (including occipital, temporal and parietal cortex), right inferior frontal sulcus, and left post-central motor cortex (the familywise error rate was controlled at the voxel-level $p < 0.001$). **B)** To isolate voxels most strongly associated with the visual feature representations, we raised the statistical threshold, resulting in the ROI illustrated in yellow. **C)** 5/4 Dataset Binary Feature Decoding. Red dots indicate scores from individual participants. **D)** SHJ Binary Feature Decoding. The same ROI ("B") was used in both datasets.

Effects Associated with Conceptual Knowledge

5/4 Dataset

First, we confirmed that each stimulus feature could be decoded significantly above chance from the ROI illustrated in Figure 3.B (yellow) [2]. The analyses were performed in native anatomical space of each participant using the cross-validated SVC analysis described above (but setting the C parameter to 1 instead of 0.1, which

[2]Note, that despite the circularity of testing voxels identified via the searchlight analysis, subsequent testing at the ROI-level has been recommended, in order to confirm that information exists not only at the level of the ROI, but also at the level of the ROI (Etzel et al., 2013).

was chosen for the searchlight analysis to improve computational efficiency). Each feature could be decoded at rates significantly above chance (Figure 3.C; shape: $M = 0.60$, $SE = 0.02$, $t(19) = 5.78$, $p < 0.001$, size: $M = 0.70$, $SE = 0.02$, $t(19) = 9.29$, $p < 0.001$, color: $M = 0.54$, $SE = 0.01$, $t(19) = 3.64$, $p = 0.002$, position: $M = 0.91$, $SE = 0.02$, $t(19) = 25.96$, $p < 0.001$).

Next, we investigated whether the discriminability of the individual perceptual dimensions covaried with the GCM attentional parameters. To do so, we fit a mixed-effects linear regression analysis (as implemented in the lme4 package for R) using restricted maximum likelihood (ReML). We included fixed-effects terms for the intercept, the attentional weights, and each visual dimension (e.g., "color"). We also included random effects terms (which were free to vary between participants) for the intercept and the attention weight parameters. This allowed us to control for baseline differences in decoding accuracy between participants, and for shared (group-level) differences in decoding accuracy between visual dimensions. We used the Kenward-Roger approximation (Kenward & Roger, 1997) to estimate degrees of freedom (reported below), and used single-sample t-tests to calculate p-values for each coefficient (using the pbkrtest package for R; Halekoh & Højsgaard, 2014)[3]. We computed 95% confidence intervals using bootstrap resampling (1000 simulations). The results indicated that the discriminability of each stimulus dimension positively covaried with the behaviorally-derived GCM parameters ($b = 0.08$, 95% CI = [0.01, 0.16], $SE = 0.04$, $t(28.71) = 2.26$, $p = 0.032$), indicating that the discriminability of these representations reflected their importance in the category structure.

To investigate the sensitivity of OTC feature representations to individual differences in GCM attentional weights, we permuted the weights associated with each participant (holding assignment of category dimension[4] constant between participants), and repeated the analysis 10,000 times. The unpermuted beta coefficient ($b = 0.08$) was significantly greater than those composing the null distribution ($P = 0.994$), indicating that the discriminability of the OTC representations was sensitive to between-subject differences in the attentional weights. This could reflect idiosyncratic differences in behavioral strategy, and/or greater attention devoted to salient perceptual dimensions. To investigate whether visual salience may have influenced attention, we conducted a repeated measures ANOVA. There was no significant relationship between the visual features and the attentional parameters ($F(3,57) = 0.68$, $p = 0.56$). A Bayesian repeated-measures ANOVA (Rouder et al., 2017), additionally indicated that the null model was 4.65 times more likely than the alternative hypothesis. These results provide evidence that the observed effects were not driven by visual characteristics of the stimulus features.

[3]This provides a more conservative test than the likelihood ratio test (fit using maximum-likelihood) or the Wald approximation (Luke, 2016).

[4]I.e., the dimensions depicted in Table 1.

After finding the attentional effect, we the compared the performance of the MPM to that of the GCM. We predicted that the GCM would perform better than the MPM, as this pattern was observed in the original study (Mack et al., 2013). The results confirmed this prediction. While the effect of selective attention in MPM on visual representation was in the predicted direction, in contrast to GCM it was not significant (b = 0.03, 95% CI = [-0.03, 0.09], SE = 0.03, t(34.69) = 0.964, p = 0.342). The AIC value associated with the GCM model (-202.8) was smaller than that of the prototype model (-195.0) indicating that the attentional weights from the GCM more accurately predicted the strength of occipitotemporal representations than those of the prototype model.

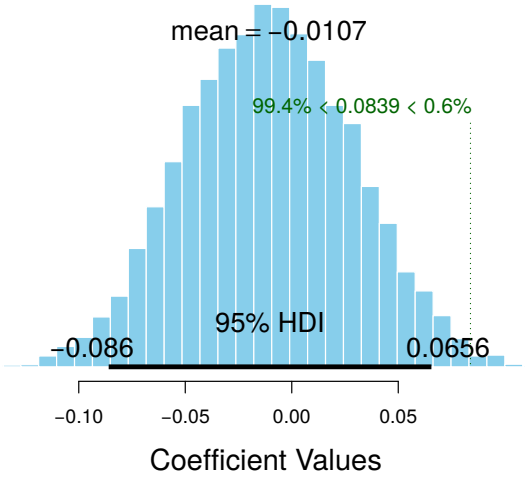

Figure 4: **Individual Differences in the 5/4 Dataset.** To investigate whether occipitotemporal representations of stimulus features reflect individual differences in conceptual knowledge, we conducted a permutation test. This involved repeatedly performing a mixed-effects regression analysis to predict the decoding accuracy of each visual feature based on the attentional weights from the GCM (see text for description of full regression model). To estimate the null distribution (M = -0.01; **blue**), the attentional weights were shuffled between participants while maintaining the mapping between abstract dimensions (i.e., the dimensions depicted in table 1). The true regression coefficient for the attentional weights estimated from unshuffled data (b = 0.08; **vertical green line**) was greater than 99.4% of those composing the null distribution. **HDI**: Highest Density Interval.

SHJ Dataset

First, we confirmed that each stimulus feature could be decoded significantly above chance from the ROI illustrated in Figure 3.B (yellow). Using a four-fold, leave-one-run out cross-validation strategy, we used a linear support vector classifier (C=1) to decode each visual feature across all runs (including both early and late learning epochs), retaining only estimates for the last two runs (which corresponded to the late-learning phase in which behavior had largely stabilized). This four-fold cross-validation strategy yielded better decoding accuracy than a two-fold approach based on only the last two runs. This improvement reflects the increased amount of training data available in the 4-fold approach, and suggests that the multivariate patterns reflecting

the individual visual features were stable across learning. Each feature could be decoded at rates significantly above chance (Figure 3.D; antennae: $M = 0.57$, $t(20) = 3.82$, $p = 0.001$; mandibles: $M = 0.56$, $t(20) = 3.22$, $p = 0.004$; legs: $M = 0.58$, $t(20) = 4.17$, $p < 0.001$).

Next, we investigated whether the decoding accuracy associated with the features covaried with SUSTAIN's attentional parameters. To do so, we used a mixed-effects linear regression analysis to predict decoding accuracy from attention weight, visual dimension, run and rule. The model included fixed-effects parameters for these four variables, and random-effects parameters for the intercept, attention weight, and run (which were free to vary by participant). This allowed us to control for differences in decoding accuracy across visual dimensions and participants (as with the model used for the 5/4 dataset), while additionally controlling for effects of rule and idiosyncratic differences in behavioral performance during the last two runs. Mirroring the findings from the 5/4 dataset, we found that the discriminability of these patterns during the late learning phase positively covaried with the attention parameters derived from SUSTAIN ($b = 0.09$, 95% CI = [0.004, 0.17], $SE = 0.04$, $t(61) = 2.13$, $p = 0.038$).

To investigate the sensitivity of OTC feature representations to individual differences in SUSTAIN's attentional parameters, we conducted a permutation test in which we shuffled the attentional weight parameters between participants 10,000 times (preserving the mapping between rule and abstract feature). The slope parameter associated with the unpermuted data was significantly greater than the permuted distribution ($P = 0.979$), suggesting that the visual feature representations were sensitive to idiosyncratic differences in attentional weights. A repeated measures ANOVA indicated that the visual dimensions did not influence the attentional parameters ($F(2,44) = 1.27$, $p = 0.291$). A Bayesian repeated measures ANOVA additionally indicated that the null model was 1.98 times more likely than the alternative hypothesis, providing evidence that the attentional weights were not influenced by visual properties of the stimulus features.

Discussion

Although differing substantially in how concepts are represented (e.g., as exemplars, prototypes, or clusters), formal categorization theories (e.g., Love et al., 2004; Kruschke, 1992; Nosofsky, 1986) share a similar conception of selective attention. In these models, conceptual knowledge “warps” psychological space (Figure 1, and equations 1 & 4 in appendix B), such that differences along behaviorally-relevant dimensions are accentuated, and differences along irrelevant dimensions are down-weighted. In two datasets (Mack et al., 2013, 2016), we evaluated the neurobiological plausibility of this idea by investigating whether occipitotemporal stimulus feature representations covaried with attentional parameters from formal categorization theory. We found that this

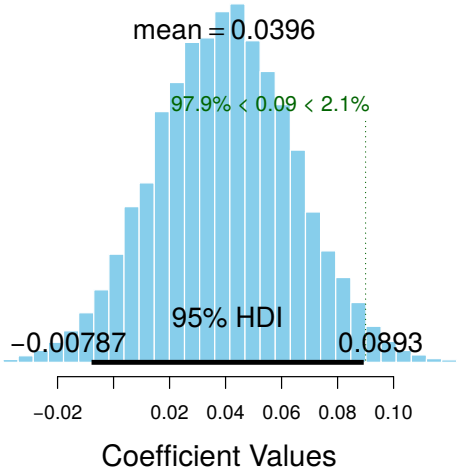

Figure 5: **Individual Differences in the SHJ Dataset**. A similar permutation test to that used for the 5/4 dataset was performed. To estimate the null distribution (M = 0.04; **blue**), the attentional weights were shuffled between participants while maintaining the mapping between rule, run and abstract dimension (i.e., the dimensions depicted in table 2). The true regression coefficient for the attentional weights estimated from unshuffled data (b = 0.09; **vertical green line**) was greater than 97.9% of those composing the null distribution. **HDI**: Highest Density Interval.

effect was not only apparent at the group-level, but was sufficiently sensitive to reflect individual differences in conceptual knowledge. This provides particularly strong evidence that occipitotemporal representational space shares fundamental characteristics with that formalized by classic categorization theory.

Our analyses of these two datasets extended the findings of the original authors. The original contributions considered the same cognitive models that we have here and evaluated correspondence of these models' attention-weighted representations to patterns of brain activity. Those analyses were a joint evaluation of both category representation (e.g., exemplar, prototype, or clusters) and attentional influence. We extended these results by directly evaluating the validity of the attention parameters. In particular, we evaluated whether these parameters related to how decodable these visual features were in visual areas of the human brain. Moreover, we found support that individuals' variations in attention parameters (fit to behavior) was reflected in individual differences in brain activity. Previously, it was unclear whether these fitted parameters were valid at the neural level, in their own right.

We are not the first to link brain and behavior via latent model parameters. In the perceptual decision-making literature, for instance, several groups have fit the drift diffusion model (Ratcliff, 1978) to behavioral data, and identified regions of the brain where the BOLD response reflects variation in its drift rate, bias, and threshold parameters (e.g., White et al., 2014; Mulder et al., 2012; Forstmann et al., 2008). As in the present study, these studies demonstrated that individual differences in behavioral strategy are reflected in the brain.

Instead of linking latent model parameters to univariate BOLD amplitude, however, we used MVPA to link latent parameters to multivariate representations of individual stimulus features. This allowed us to investigate how idiosyncratic differences in selective attention influence occipitotemporal representations. Several studies have demonstrated that occipitotemporal representations are modulated by selective attention (e.g., Reynolds & Chelazzi, 2004; Moran & Desimone, 1985; Zhang et al., 2011; Reynolds et al., 2000), and by learned conceptual knowledge (e.g., Li et al., 2007; Folstein et al., 2013; Dieciuc et al., 2017; Sigala & Logothetis, 2002; Sigala, 2004). These studies, however, have relied on contrastive approaches, in which representations of relevant stimulus dimensions were compared to those of irrelevant dimensions. By linking behavior and brain by testing formal categorization models, we demonstrate that these attentional effects are sufficiently sensitive to reflect fine-grained, idiosyncratic differences in conceptual knowledge.

Interestingly, although category training can improve perceptual discriminability of relevant stimulus features when stimuli consist of perceptually-separable features (Garner, 1976), this may not occur for integral dimensions (Op de Beeck et al., 2003) or for stimuli defined according to "blended" stimulus morphspaces (Folstein et al., 2013). More work is needed to better understand how attention influences occipitotemporal representations for such stimuli. One possibility is that selective attention does not warp *perceptual* representations of integral stimulus dimensions, but might operate on abstract cognitive or "decisional" representations in higher-order cortex (Jiang et al., 2007; Nosofsky, 1987).

Conclusions

Category training is known to induce changes in both perceptual (Folstein et al., 2012; Op de Beeck et al., 2003; Goldstone et al., 1996; Goldstone, 1994; Gureckis & Goldstone, 2008) and neural sensitivity (e.g., Li et al., 2007; Folstein et al., 2013, 2015; Dieciuc et al., 2017; Sigala, 2004; Sigala & Logothetis, 2002; Sigala, 2004). In two datasets, we demonstrate that occipitotemporal stimulus representations covary with the attentional parameters derived from formal categorization theory. This effect was sufficiently sensitive to reflect individual differences in conceptual knowledge, which implies that these occipitotemporal representations are embedded within a space closely resembling that predicted by formal categorization theory (e.g., Nosofsky, 1986; Love et al., 2004; Kruschke, 1992).

Appendix A

Size	x	y	z	*t*	BA	Region
23972	14	-74	4	12.717	18	Calcarine_R
	28	-62	54	10.5	7	Parietal_Sup_R
	-54	-18	44	7.256	3	Postcentral_L
233	50	24	26	7.447	48	Frontal_Inf_Tri_R
22	-46	12	44	6.621	9	Frontal_Mid_L

Table 3: 5/4 Dataset Binary Feature Decoding: Searchlight Results. The familywise error rate was controlled at the voxel level ($p < 0.001$)

Appendix B: Computational Models and Attentional Parameters

For the first dataset (Mack et al., 2013), we considered the Generalized Context Model (GCM; Nosofsky, 1986), which posits that conceptual knowledge consists of memory for individual exemplars. For the SHJ experiment (Mack et al., 2016), we considered the attentional parameters from SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network; Love et al., 2004). Details about the models can be found in the original papers. Here, we provide a brief overview of each.

GCM

In the Generalized Context Model (GCM; Nosofsky, 1987), the psychological distance, d between stimuli i and j can be calculated as the attentionally-weighted sum of their unsigned differences across dimensions, k:

$$d_{ij} = \sum_k \left[w_k |x_{ik} - x_{jk}|^r \right]^{1/r}, \tag{1}$$

where w indicate the attentional parameters assigned to each dimension. The r parameter is set to 1 (city-block distance) for perceptually separable stimulus dimensions (as in the 5/4 dataset), and r is set to 2 (Euclidean distance) for integral dimensions (Garner, 1976). Similarity is an exponentially-decaying function of psychological distance:

$$s_{ij} = d^{-cd_{ij}}, \tag{2}$$

where the shape of the similarity gradient is influenced by the sensitivity parameter, c. The probability of choosing category "A", given stimulus, i, is given by the choice rule:

$$P(A|i) = \frac{\left(\sum_{a \in A} s_{ia} \right)^{\gamma}}{\left(\sum_{a \in A} s_{ia} \right)^{\gamma} + \left(\sum_{b \in B} s_{ib} \right)^{\gamma}}, \tag{3}$$

where γ governs the degree of deterministic responding.

SUSTAIN

SUSTAIN is a semi-supervised clustering model, which incrementally learns to solve categorization problems by first applying simple solutions, and then increasing complexity as required. Through experience, the model can learn to group similar items into common clusters, and can make inferences about novel stimuli based on its perceptual similarity to existing clusters (i.e., based on perceptual similarity, clusters compete to predict

latent stimulus attributes). When unexpected feedback is received, the model can also learn in a supervised fashion by creating a new cluster to represent the novel stimulus.

In SUSTAIN, all clusters contain receptive fields (RF's) for each stimulus dimension. As new stimuli are added to the cluster, the model learns by adjusting the position of each RF to best match the cluster's expectation for novel stimuli. As the RF is an exponential function, a cluster's activation, α, decreases exponentially with distance from its preferred value:

$$\alpha(\mu) = \lambda e^{-\lambda\mu}, \tag{4}$$

where μ represents the distance of the stimulus dimension value from the cluster's preferred stimulus dimension value, and where λ represents the tuning (or width) of the RF. The λ parameters are specific to dimensions, but are shared across dimensions, and so, like the attentional parameters in the GCM, the λ parameters in SUSTAIN modulate the influence of each stimulus dimension on the overall decision outcome.

The overall activation of a cluster, H, involves consideration of each dimension, k:

$$H = \frac{\sum_k (\lambda_k)^\gamma e^{-\lambda_k \mu_k}}{\sum_k (\lambda_k)^\gamma}, \tag{5}$$

where the γ parameter (which is always non-negative) modulates the influence of the λ parameters on the choice outcome. When γ is large, attended dimensions (which are associated with large λ values, and narrow RF's), dominate the activation function (eq. 5); when γ is zero, the λ parameters are ignored, and all dimensions exert equal influence on the choice.

SUSTAIN was fit to the SHJ dataset in a supervised fashion, using the same trial order experienced by the participants; it was also fit across rule-switches, such that learning from one task was carried over to the next. Thus, SUSTAIN was capable of reflecting learning, as well as carry-over effects associated with previously learned rules.

References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Avants, B., Tustison, N., & Johnson, H. (2009). Advanced Normalization Tools (ANTS). *Insight Journal*, (pp. 1–35).

Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science (New York, N.Y.)*, *344*, 424–7.

Bichot, N. P., Heard, M. T., Degennaro, E. M., & Desimone, R. (2015). A source for feature-based attention in the prefrontal cortex. *Neuron*, *88*, 1–13.

Buffalo, E. A., Fries, P., Landman, R., Liang, H., & Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences*, *107*, 361–365.

Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron*, *76*, 838–846.

Dieciuc, M., Roque, N. A., & Folstein, J. R. (2017). Changing similarity: Stable and flexible modulations of psychological dimensions. *Brain Research*, *1670*, 208–219.

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, *78C*, 261–269.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: Not all morphspaces stretch alike. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*, 807–20.

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

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

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 17538–42.

Garner, W. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology*, *8*, 98–123.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology. General*, *123*, 178–200.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann, & C. Olson (Eds.), *Perceptual organization in vision: Behavioral and neural perspectives.* chapter 7. (pp. 233–278). New Jersey: Lawrence Erlbaum Associates.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In *Proceedings of the eighteenth annual conference of the Cognitive Science Society* (pp. 243–248).

Gregoriou, G. G., Rossi, A. F., Ungerleider, L. G., & Desimone, R. (2014). Lesions of prefrontal cortex reduce attentional modulation of neuronal responses and synchrony in V4. *Nature Neuroscience*, *17*, 1003–11.

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, (pp. 1876–1881).

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models - The R package pbkrtest. *Journal of Statistical Software*, *59*, 1–32.

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.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3863–8.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, .

Li, S., Ostwald, D., Giese, M., & Kourtzi, Z. (2007). Flexible coding for categorical decisions in the human brain. *The Journal of Neuroscience*, *27*, 12321–30.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–32.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, *77*, 24–42.

Luke, S. G. (2016). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*, 1494–1502.

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

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

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

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 275–292.

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*, 103–18.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science (New York, N.Y.)*, *229*, 782–784.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology*, *70*, 909–919.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*, 2335–2343.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*, 2636–43.

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.

Nosofsky, R. M. R. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *13*, 87–108.

Nosofsky, R. R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *115*, 39–57.

Op de Beeck, H., 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.

Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 59–64.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poort, J., Khan, A. G., Pachitariu, M., Nemri, A., Orsolic, I., Krupic, J., Bauza, M., Sahani, M., Keller, G. B., Mrsic-Flogel, T. D., & Hofer, S. B. (2015). Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron*, *86*, 1478–1490.

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

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611–647.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*, 703–714.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*, 304–321.

Schwedhelm, P., Baldauf, D., & Treue, S. (2017). Electrical stimulation of macaque lateral prefrontal cortex modulates oculomotor behavior indicative of a disruption of top-down attention. *Scientific Reports*, *7*, 1–10.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Shepard, R. N., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*.

Sigala, N. (2004). Visual categorization and the inferior temporal cortex. *Behavioural Brain Research*, *149*, 1–7.

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

Townsend, J. T., & Ashby, F. G. (1982). Experimental test of contemporary mathematical models of visual letter recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *8*, 834–854.

Tsotsos, J. K. (2011). *A Computational Perspective on Visual Attention*. MIT Press.

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.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327:352.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*, 953–6.

White, C. N., Congdon, E., Mumford, J. A., Karlsgodt, K. H., Sabb, F. W., Freimer, N. B., London, E. D., Cannon, T. D., Bilder, R. M., & Poldrack, R. A. (2014). Decomposing decision components in the stop-signal task: A model-based approach to individual differences in inhibitory control. *Journal of Cognitive Neuroscience*, *26*, 1601–14.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1160–1173.

Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., & Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *108*, 8850–8855.