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 increase the perceptual discriminability of relevant stimulus information. To account for this effect, popular contemporary cognitive models of categorization typically include of attentional parameters, which provide information about the importance of each stimulus dimension in decision-making. The effect of these parameters on psychological representation is 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). In sensory and association cortex, representations of stimulus features are known to covary with their behavioral relevance. Although this implies that neural representational space might closely resemble that hypothesized by formal categorization theory, to date, attentional effects in the brain have been demonstrated through powerful experimental manipulations (e.g., contrasts between relevant and irrelevant features). This approach sidesteps the role of idiosyncratic conceptual knowledge in guiding attention to useful information sources. To bridge this divide, we used formal categorization models, which were fit to behavioral data, to make inferences about the concepts and strategies used by individual participants during decision-making. We found that when greater attentional weight was devoted to a particular visual feature (e.g., "color"), its value (e.g., "red") was more accurately decoded from occipitotemporal cortex. We additionally found that this effect was sufficiently sensitive to reflect individual differences in conceptual knowledge. The results indicate that occipitotemporal stimulus representations are embedded within a space closely resembling that proposed by classic categorization models.

Keywords: Concepts, Selective Attention, Occipitotemporal Cortex

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

---

*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 evolving conceptual knowledge (Tversky, 1977; Goldstone, 2003). In formal theories of categorization, one way to account for this is through inclusion of attentional parameters, which reflect the influence of each dimension on the final choice (e.g., Kruschke, 1992; Nosofsky, 1986; Love et al., 2004). These parameters are often described as "warping" multidimensional 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 these attentional parameters predict the strength of neural stimulus feature representations. Importantly, we link individual differences in conceptual knowledge (as revealed by model fits of attentional parameters) to individual differences in neural representation (as revealed by decoding stimulus features in fMRI data). In doing so, we bridge behavioral and neural levels of analysis at the individual level using cognitive models.
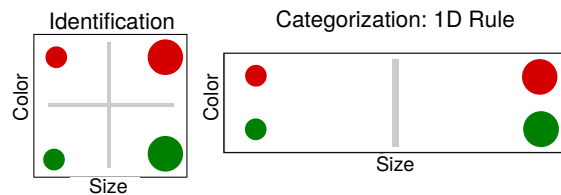

Figure 1: **Example: Attention Influences Psychological Space**. **Left:** In an object identification task, both psychological dimensions should receive equivalent attention, as they are equally relevant. **Right:** In a one-dimensional rule-based categorization task, only a single dimension is relevant (in this example, 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").

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 distinct 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, as prototypes, or as clusters that flexibly reflect environmental structure; Love et al., 2004; Nosofsky, 1987; Nosofsky & Zaki, 2002; Smith & Minda, 1998; Minda & Smith, 2002; Zaki et al., 2003), they assume that categorization involves learning to distribute attention across stimulus features so as to

2

optimize behavioral performance. Although they differ in their mathematical details, these models also propose that *endogenous* (i.e., "top-down") attentional control (Miller & Cohen, 2001; Tsotsos, 2011) can modulate the influence of the *exogenous* (or perceptual) stimulus dimensions on the behavioral choice. The attentional parameters play a key role in allowing the models to capture patterns of human generalization across different goals and rules. They also predict human eye-movements during category decision-making (e.g., Rehder & Hoffman, 2005a,b), and are therefore thought to reflect the strategies used by decision-makers to actively-sample information from the external world.

In the brain, effects of endogenous attention have been observed across the visual cortical hierarchy (Buffalo et al., 2010; Jehee et al., 2011; Kamitani & Tong, 2005, 2006; Luck et al., 1997; Motter, 1993). A general finding is that when attention is devoted to a specific visual feature, its neural representation is more accurately decoded. For instance, in human fMRI, when multiple visual gratings are concurrently presented, representations of attended orientations in areas V1-V4 are more easily decoded than those that are unattended (Kamitani & Tong, 2005; Jehee et al., 2011). Similarly, when random dot stimuli move in multiple directions, representations of attended motion directions in area MT+ are more easily decoded than those that are unattended (Kamitani & Tong, 2006). Whereas these studies have relied on explicit cues to guide attention to relevant aspects of the stimulus array, in real-world environments, decision-makers must typically rely on knowledge gained through past experience in order to selectively attend to relevant information sources.

Categorization tasks mirror this aspect of real-world environments; decision-makers must rely on learned conceptual knowledge in order to selectively attend to relevant stimulus dimensions. Several studies have investigated whether neural representations of exogenous information sources are modulated by learned conceptual knowledge (e.g., Li et al., 2007; Folstein et al., 2013; Sigala & Logothetis, 2002). Sigala & Logothetis (2002), for instance, trained macaques to categorize abstract images, which varied according to four stimulus dimensions. Neural representations of the two behaviorally-relevant stimulus dimensions (i.e., the dimensions that reliably predicted the correct response) in the inferior temporal lobe were enhanced relative to those of the irrelevant dimensions. Using fMRI with human participants, Li et al. (2007) investigated whether neural representations of stimulus motion and shape were influenced by their relevance to the active categorization rule. Using multivariate pattern analysis (MVPA), they similarly found that representations of these stimulus dimensions reflected their relevance to the active rule.

Across studies involving explicit attentional cues and categorization studies involving learned conceptual knowledge, a known effect is that occipitotemporal representations of behaviorally-relevant information sources are enhanced relative to those that are irrelevant (but this may not hold for integral stimulus dimensions; Garner,

1976). These effects are compelling, as they imply that occipitotemporal representational space may closely resemble that conceptualized by classic cognitive theory (e.g., Love et al., 2004; Kruschke, 1992; Nosofsky, 1986). Specifically, these representations may expand and contract, along axes defined by perceptually-separable stimulus dimensions (Garner, 1976), in ways that closely reflect the idiosyncratic concepts and strategies used by individual participants during decision-making.

Previous studies have relied on contrastive analyses in which neural representations of attended stimulus dimensions are compared to those of unattended dimensions. Although statistically-powerful, this approach defines selective attention in terms of the experimental paradigm (but see O'Bryan et al., 2018), and therefore sidesteps effects associated with individual differences in conceptual knowledge (e.g., Craig & Lewandowsky, 2012; Little & McDaniel, 2015; McDaniel et al., 2014; Raijmakers et al., 2014). These effects can be substantial, particularly for ill-defined categorization-problems (such as the 5/4 categorization task) that are common in every-day life (Hedge et al., 2017; Johansen & Palmeri, 2002). Here, we seek to address this divide by combining model-based fMRI (Turner et al., 2017; Palmeri et al., 2017) with multivariate pattern analyses. This allowed us to abstract away from individual differences in neural topography (Kriegeskorte & Kievit, 2013; Haxby et al., 2001; Haynes, 2015), to investigate whether neural stimulus representations reflect individual differences in conceptual knowledge. Specifically, we sought to investigate whether the attentional parameters derived from formal categorization models predict contortions of occipitotemporal representational space during decision-making.

We investigated this hypothesis using two publicly available datasets (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 (Nosofsky, 1987; Nosofsky & Zaki, 2002) to the behavioral data, and used them to compare exemplar and prototype accounts of occipitotemporal representation. Using representational similarity analysis (Kriegeskorte et al., 2008), Mack et al. (2013) additionally identified regions of the brain (lateral occipital cortex, parietal cortex, inferior frontal gyrus, and insular cortex) sensitive to the attentionally-modulated pairwise similarities between stimuli. Although these results (particularly those in lateral occipital cortex) imply that neural representations of the individual stimulus features might be modulated by selective attention, in principle, this could also reflect modulation within an abstract representational space where stimulus features are not individually represented. For instance, while visual cortex reflects sensory input (and is known to represent individual stimulus dimensions), prefrontal cortex can flexibly represent conjunctions of features, abstract rules, and category boundaries in a

4

goal-directed manner. Representations in parietal cortex display intermediate characteristics, as they can reflect both sensory and decisional factors (Brincat et al., 2017; Jiang et al., 2007; Li et al., 2007).

In the second dataset (Mack et al., 2016), participants learned, while scanning, to categorize images of insects that varied according to three binary perceptual dimensions (Figure 2.B), according to type I, type II and type VI problems described by Shepard et al. (1961).[1] Importantly, although the same stimuli were included in each task, the degree to which each of the features predicted the correct choice differed between rules. 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. Using representational similarity analysis, they found that SUSTAIN successfully predicted the pairwise similarities between hippocampal stimulus representations across rule-switches. This suggests that hippocampal representations are updated according to goal-directed attentional selection of stimulus features.

Methods

Description of Datasets

In both experiments, participants categorized stimuli that were characterized by multiple perceptually-separable stimulus dimensions. As the mapping of perceptual attributes to their role in each category structure was randomized for each participant, it is possible to differentiate effects associated with intrinsic perceptual stimulus attributes from effects of behavioral relevance. For example, while color strongly predicted the correct category choice for some participants, it provided unreliable informative for others. In both experiments, participants were not instructed as to which cues were informative, and learned to perform each task through trial-and-error.

We used the GCM for the first dataset (the winning model from Mack et al. 2013), as participants learned how to perform the categorization task prior to scanning. We used SUSTAIN for the second dataset, as it learns on a trial-by-trial basis, and participants learned to perform each task during scanning. SUSTAIN was additionally fit in such a way that the learning of one task carried over to the next. Importantly, although the GCM and SUSTAIN differ in how stimuli are represented in memory (i.e., as exemplars or clusters), they similarly posit that attention "contorts" psychological space, as illustrated in Figure 1. Thus, these studies and models provide a good test of whether attention weights in successful cognitive models are plausible at both

---

[1]In their paper, Mack et al. (2016) focus on effects associated with the type I and type II rules.

A) 5/4 Experiment

Color:	red	green
Size:	large	small
Shape:	circle	triangle
Position:	left	right

B) SHJ Experiment

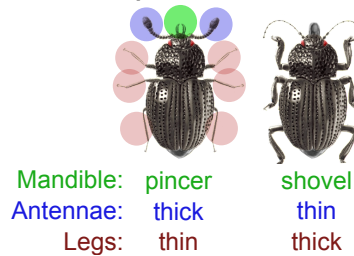| Mandible:	pincer	shovel
Antennae:	thick	thin
Legs:	thin	thick

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 2 and 3) was randomized for each participant.

behavioral and neural levels of analysis.

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 to categorize the stimuli according to the "5/4" categorization task (Medin & Schaffer, 1978) through trial-and-error. During this training session, participants were shown only the first nine stimuli shown in Table 2 (i.e., five category "A" members: A1-A5 and four category "B" members: B1-B4), and experienced 20 repetitions of each stimulus. During the anatomical scan, they additionally performed a "refresher" task, involving four additional repetitions on each training item. Each training trial involved a 3.5 second stimulus presentation period in which participants made a button press. Following the button press, a fixation cross was shown for 0.5 seconds, and feedback was then presented for 3.5 seconds. Feedback included information about the correct category, and about whether the response was correct or incorrect. During scanning, participants were required to categorize not only the training items, but also 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

6

130   stimulus presentations were randomized for each participant.

The SHJ Dataset

132   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:

134 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

136 to categorize the stimuli according to the type I, type II, and type VI problems described by Shepard et al. (1961). In the type I problem, the optimal strategy required attending to a single stimulus dimension (e.g.,

138 "legs") that perfectly predicted the category label, while ignoring the other two dimensions. In the type II problem, the optimal strategy was a logical XOR rule, in which two stimulus features had to be considered

140 together. In the type VI 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

142 and exogenous factors, the irrelevant feature in the type II rule was used as a relevant feature of the type I problem for each participant.

144   Each problem was performed across four scanner runs. While all of the participants learned to perform the type VI problem first, the order of the type I and type II problems was then counterbalanced across participants.

146 Each trial consisted of a 3.5 second stimulus presentation period, a jittered 0.5-4.5 second fixation period, and feedback. Feedback was presented for 2 seconds and consisted of an image of the presented insect, as well

148 as text indicating whether the response was correct or incorrect. Each trial was separated by jittered intertrial interval (4-8 seconds), which consisting of a fixation cross. Each run included four presentations of each of the

150 eight stimuli.

  For consistency across datasets, we used the group-derived region of interest (ROI) used in "5/4" dataset

152 (Figure 3.B), and performed a similar analysis. As participants in the SHJ experiment learned to perform the type I, type II and type VI problems during scanning, we mirrored the strategy used by the original authors, and

154 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

156 during this late learning phase, in which behavior had largely stabilized.

  SUSTAIN was initialized with no clusters, and with equivalent weights assigned to each stimulus dimension.

158 Its learning parameters were first fit to the learning performance of each participant using a maximum-likelihood genetic algorithm procedure. The model was fit in such a way that, after learning one problem, the model state

160 was used as the initial state for the subsequent problem. In this way, the model was fit under the assumption that

learning of one task would influence later behavior. Once the learning parameters of the model were optimized,

162 they were fixed, and the attentional parameters were extracted from the second two runs of each task (in which

learning had largely stabilized). This yielded distinct sets of attentional parameters for each participant and each

164 task. More information about the model can be found in appendix C.

Image Processing

166   Preprocessing included motion correction, and coregistration of the anatomical images to the mean of the

functional images (using Statistical Parametric Mapping (SPM), version 6470). All MVPA analyses were per-

168 formed in native space without smoothing. For group-level analyses, the statistical maps from each participant

were warped to Montreal Neurological Institute (MNI) atlas space using Advanced Normalization Tools (ANTs;

170 Avants et al., 2009), and then smoothed with a 6mm full-width at half maximum Gaussian kernel. The ROI de-

rived from group-level analyses were transformed back into each participants native space for ROI-level analy-

172 ses. 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

174 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.

176 **Results**

Representations of Individual Visual Features

178   To identify regions most strongly representing the stimulus features, we performed a cross-validated search-

light analysis (sphere radius = 10mm.; Kriegeskorte et al., 2006)[2] in which we decoded each of the four visual

180 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 con-

182 junction 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

184 with the held-out neuroimaging data.

    After centering each of the resultant statistical maps at chance (50% for each visual feature), we created a

186 single map for each participant, which reflected the average, above chance, decoding accuracy across features.

We then normalized each map to to MNI space and, in order to identify regions supporting above chance feature

---

[2]This involves moving an imaginary sphere throughout the brain; repeatedly investigating how well the voxels within the sphere can decode a variable of interest.

188 decoding, performed a group-level permutation test. This involved randomly flipping the sign of the statistical

maps 10,000 times (using the randomise function from the Oxford Centre for Functional MRI of the Brain

190 Software Library (FSL); 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 post-central motor cortex, as

192 well as widespread visual and association cortex, extending dorsally from occipital pole to the bilateral superior

extrastriate cortex and bilateral intraparietal sulcus (IPS), and ventrally into the bilateral lingual gyrus (Table

194 1). 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 individual stimulus features. This removed

196 voxels belonging to the bilateral inferior occipital cortex, left lingual gyrus, bilateral intraparietal sulcus, and

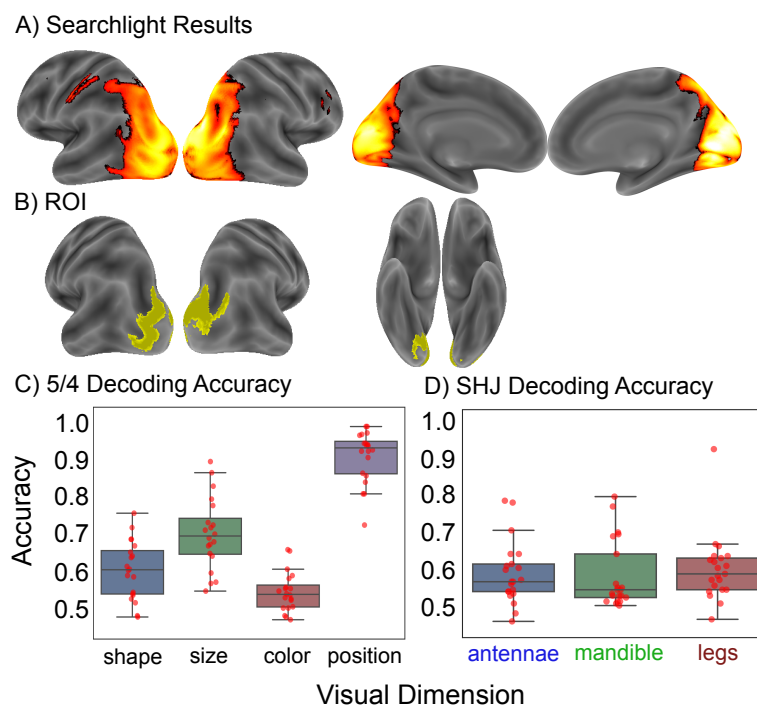bilateral precuneus. The resultant ROI is illustrated in Figure 3.B.

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 family-wise error rate was controlled at the voxel-level $p < 0.001$). **B)** To isolate voxels most strongly representing the stimulus features, 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. Although estimating effect sizes on voxels selected through non-orthogonal criteria is circular, testing significance at the ROI-level has been recommended to confirm that information exists, not only at the level of the searchlight sphere, but also at the level of the ROI (Etzel et al., 2013). This analysis also allows us to illustrate the individual feature decoding accuracies for each participant (Figure 3.C). The analyses were performed in the 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 was chosen for the searchlight analysis to improve computational efficiency).[3] Each feature could be decoded at rates significantly above chance (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 decoding accuracy 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)[4]. We computed 95% confidence intervals using bootstrap resampling (1000 simulations). The decoding accuracy 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 decoding accuracy of these representations reflected their importance during decision-making.

To investigate the sensitivity of occipitotemporal feature representations to individual differences in GCM attentional weights, we conducted a permutation test. This involved shuffling the attentional weight parameters between participants (i.e., swapping the weights derived from one participant with those derived from another),

---

[3]The C parameter modulates the penalty associated with training error. With large values, the classifier will choose a small-margin hyperplane, and training accuracy will be high. With smaller values, out-of-sample performance is often improved, but more training samples may be misclassified. C=1 is a common default setting for fMRI.

[4]This provides a more conservative test than the likelihood ratio test or the Wald approximation (Luke, 2016).

226 and repeating the regression analysis (described above) 10,000 times. On each permutation, the correspondence for category dimensions (i.e., the dimensions depicted in Table 2, as opposed to the stimulus dimensions illus-

228 trated in Figure 1) was preserved, such that the dimensional weights derived from the behavior of one participant were assigned to the same dimensions, but to a different participant.

230 The unpermuted beta coefficient ($b = 0.08$) was significantly greater than those composing the null distribution ($P = 0.994$), indicating that the decoding accuracy of the occipitotemporal representations was sensitive to

232 between-subject differences in the attentional weights. This could reflect idiosyncratic differences in behavioral strategy, and/or effects associated with perceptual saliency. Therefore, to investigate whether visual salience

234 may have influenced attention, we conducted a repeated measures ANOVA for the perceptual features. There was no significant relationship between these visual features and the attentional parameters ($F(3,57) = 0.68$, $p =$

236 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

238 effects were not driven by visual characteristics of the stimulus features.

SHJ Dataset

240 First, we confirmed that each stimulus feature could be decoded significantly above chance from the ROI illustrated in Figure 3.B. Using a four-fold, leave-one-run out cross-validation strategy, we used a linear support

242 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

244 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

246 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

248 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: $M = 0.58$, $t(20) = 4.17$, $p < 0.001$).

250 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

252 from attention weight, visual dimension, run and rule. As described in the Methods section, distinct attentional weights were derived for each subject and each rule. The decoding accuracy for each separate run was included

254 in the analysis. 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

256 control for differences in decoding accuracy across visual dimensions and participants (as with the model used

11

for the "5/4" dataset), while additionally controlling for effects of rule and idiosyncratic differences in behav-
258 ioral performance during the last two runs. Mirroring the findings from the "5/4" dataset, we found that the
decoding accuracy of these patterns positively covaried with the attention parameters derived from SUSTAIN
260 ($b = 0.09$, 95% CI = [0.004, 0.17], $SE = 0.04$, $t(61) = 2.13$, $p = 0.038$).

To investigate the sensitivity of occipitotemporal feature representations to individual differences in SUS-
262 TAIN's attentional parameters, we conducted a permutation test similar to that described above (i.e., for the
"5/4" experiment). This involved shuffling the attentional weight parameters between participants 10,000 times
264 (preserving the correspondence for both rule and abstract feature). This means that the attentional weight de-
rived from the behavior of one participant, for one particular rule and one particular category feature, was
266 assigned to the same rule and feature, but to a different participant. The slope parameter associated with the
unpermuted data ($b = 0.09$) was significantly greater than those composing the permuted null distribution (P
268 $= 0.979$), suggesting that the visual feature representations were sensitive to idiosyncratic differences in atten-
tional weights. A repeated measures ANOVA indicated that the perceptual dimensions did not influence the
270 attentional parameters ($F(2,44) = 1.27$, $p = 0.291$). A Bayesian repeated measures ANOVA additionally indi-
cated that the null model was 1.98 times more likely than the alternative hypothesis, providing evidence that the
272 attentional weights were not influenced by visual properties of the stimulus features.

Discussion

274 Although differing substantially in how concepts are represented (e.g., as exemplars, prototypes, or clus-
ters), formal categorization theories (e.g., Love et al., 2004; Kruschke, 1992; Nosofsky, 1986) tend to share a
276 similar conception of selective attention. In these models, conceptual knowledge contorts multidimensional psy-
chological space such that differences along behaviorally-relevant dimensions are accentuated, and differences
278 along irrelevant dimensions are down-weighted (Figure 1, and Equations 1 & 4 in appendix C). In two datasets
(Mack et al., 2013, 2016), we evaluated the neurobiological plausibility of this idea by investigating whether
280 occipitotemporal stimulus feature representations covaried with attention parameters derived from formal cat-
egorization models. We found that this effect was not only apparent at the group-level, but was sufficiently
282 sensitive to reflect individual differences in conceptual knowledge.

Several previous studies have demonstrated that occipitotemporal stimulus representations are modulated
284 by selective attention (e.g., Buffalo et al., 2010; Jehee et al., 2011; Kamitani & Tong, 2005, 2006; Luck et al.,
1997; Motter, 1993; Reynolds et al., 2000; Reynolds & Chelazzi, 2004) and by learned conceptual knowledge
286 (e.g., Folstein et al., 2013; Li et al., 2007; Sigala & Logothetis, 2002). These studies have relied on statistically-

12

powerful contrastive approaches, in which representations of attended stimulus dimensions are compared to those of unattended dimensions. A known effect is that neural signatures of attended stimulus dimensions are more discriminable than those that are unattended. This implies that occipitotemporal representational space might resemble that conceptualized by formal categorization theory (e.g., Nosofsky, 1986; Love et al., 2004; Kruschke, 1992). Specifically, the expansion and contraction of this space might reflect individual differences in the importance assigned to each stimulus dimension. However, as the contrastive approach defines selective attention with regards to the experimental paradigm, it is insensitive to individual differences in categorization strategy (e.g., Craig & Lewandowsky, 2012; Little & McDaniel, 2015; McDaniel et al., 2014; Raijmakers et al., 2014). Here, we link individual differences in behavior to individual differences in neural representation through consideration of the attentional parameters derived from formal categorization models.

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., Mulder et al., 2012; Forstmann et al., 2008; Purcell et al., 2010). As in the present study, several of 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 multivoxel representations of the stimulus features. This provided a precise test of the idea that selective attention contorts neural representational space.

These endogenous attentional effects are thought to arise through communication with other areas of the brain. In lateral frontal cortex, for instance, effects of endogenous attention occur earlier in time than in occipitotemporal cortex (Bichot et al., 2015; Baldauf & Desimone, 2014; Zhou & Desimone, 2011). Inactivation of these frontal regions (e.g., ventral prearcuate sulcus, or entire lateral prefrontal cortex) has also been associated wtih a reduction in the magnitude of attentional effects in occipitotemporal cortex (Bichot et al., 2015; Gregoriou et al., 2014). Interestingly, contextually-sensitive effects of endogenous attention have also been observed in the lateral geniculate nucleus (LGN), suggesting that some aspects of attention precede those in cortex (O'Connor et al., 2002; McAlonan et al., 2008; Saalmann & Kastner, 2011).

Finally, although we observed effects of selective attention across two different stimulus sets (abstract shapes in the "5/4" experiment, and insects in the SHJ experiment), and across multiple category structures (the "5/4" problem described by Medin & Schaffer (1978), and the Type I, II and VI problems described by Shepard et al. (1961)), these effects might not be apparent for all stimuli and tasks. For instance, although category training can improve perceptual discriminability of relevant stimulus features when stimuli consist of perceptually-separable

13

318 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

320 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

322 on abstract cognitive or "decisional" representations in higher-order cortex (Jiang et al., 2007; Nosofsky, 1987).

Conclusions

324 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., Dieciuc

326 et al., 2017; Folstein et al., 2013, 2015; Li et al., 2007; Sigala & Logothetis, 2002). In two datasets, we demonstrate that occipitotemporal stimulus representations covary with the attentional parameters derived from formal

328 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

330 that predicted by formal categorization theory (e.g., Nosofsky, 1986; Love et al., 2004; Kruschke, 1992).

By linking brain and behavior through the latent attentional parameters of cognitive models, we also link

332 two (somewhat) disparate literatures. In the neuroscience literature, effects of selective attention are typically examined using highly-structured decision problems, and selective attention is investigated by contrasting dif-

334 ferent aspects of the experimental design (i.e., relevant vs. irrelevant stimulus dimensions). In the cognitive categorization literature, researchers have focused on developing models that accurately account for behavioral

336 patterns of generalization across different goals and tasks. Our results indicate that these cognitive models can be used to examine effects of selective attention in the brain. This is the case, even for ill-defined decision

338 problems (such as the "5/4" task), as the models are able to successfully account for individual differences in conceptual knowledge.

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

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

Bichot, N. P., Heard, M. T., Degennaro, E. M., & Desimone, R. (2015). A source for feature-based attention in 344 the prefrontal cortex. *Neuron*, *88*, 1–13.

Brincat, S. L., Siegel, M., Nicolai, C. V., & Miller, E. K. (2017). Gradual progression from sensory to task-346 related processing in cerebral cortex, .

Buffalo, E. A., Fries, P., Landman, R., Liang, H., & Desimone, R. (2010). A backward progression of attentional 348 effects in the ventral stream. *Proceedings of the National Academy of Sciences*, *107*, 361–365.

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you 350 learn. *Quarterly Journal of Experimental Psychology*, *65*, 439–464.

Dieciuc, M., Roque, N. A., & Folstein, J. R. (2017). Changing similarity: Stable and flexible modulations of 352 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 356 all morphspaces stretch alike. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*, 807–20.

358 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.

360 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.

362 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* 364 *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*, 366 *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.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distrubuted and overlapping representations of face and objects in ventral temporal cortex. *Science*, *293*, 2425–2430.

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, *87*, 257–270.

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, (pp. 1–21).

Jehee, J. F. M., Brady, D. K., & Tong, F. (2011). Attention improves encoding of task-relevant features in the human visual cortex. *The Journal of Neuroscience*, *31*, 8210–8219.

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.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482–553.

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

394 Kamitani, Y., & Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Current Biology*, *16*, 1096–1102.

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

398 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.

400 Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in cognitive sciences*, *17*, 401–12.

402 Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

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

406 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.

408 Little, J. L., & McDaniel, M. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & cognition*, *43*, 283–97.

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

412 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.

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

416 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 418 America*, *113*, 13203–13208.

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its 420 neural implementation. *Current Biology*, *23*, 2023–2027.

17

McAlonan, K., Cavanaugh, J., & Wurtz, R. H. (2008). Guarding the gateway to cortex with attention in visual thalamus. *Nature*, *456*, 391–394.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology.*, *143, 143*, 668, 668–693.

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.

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

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

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.

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

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *5*, 1203–1209.

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.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*, 1113–1143.

Raijmakers, M. E. J., Schmittmann, V. D., & Visser, I. (2014). Costs and benefits of automatization in category learning of ill-defined rules. *Cognitive Psychology*, *69*, 1–24.

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

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41.

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *31*, 811–29.

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.

Saalmann, Y. B., & Kastner, S. (2011). Cognitive and perceptual functions of the visual thalamus. *Neuron*, *71*.

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

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

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.

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.

Zhou, H., & Desimone, R. (2011). Feature-Based attention in the frontal eye field and area V4 during visual search. *Neuron*, *70*, 1205–1217.

'

Appendix A: Searchlight Results

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 1: "5/4" Dataset Binary Feature Decoding: Searchlight Results. The familywise error rate was controlled at the voxel level ($p < 0.001$)

500 **Appendix B: Category Structures**

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

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 3: **SHJ Category Structures.** Participants learned by trial-and-error to perform the type I (a one-dimensional rule-based categorization task), type II (a two-dimensional XOR rule-based categorization task), and type VI (a three-dimensional task requiring memorization of the individual stimuli) problems during scanning. For each participant, perceptual stimulus dimensions (Figure 2) were randomly assigned to these abstract category dimensions.

Appendix C: Computational Models and Attentional Parameters

502   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

504   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

506   found in the original papers. Here, we provide a brief overview of each.

GCM

508   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:

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

510   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

512   distance) for integral dimensions (Garner, 1976). Similarity is an exponentially-decaying function of psychological distance:

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

514   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:

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

516   where γ governs the degree of deterministic responding.

SUSTAIN

518   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

520   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

522 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.

524 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

526 for novel stimuli. As the RF is an exponential function, a cluster's activation, α, decreases exponentially with distance from its preferred value:

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

528 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

530 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.

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

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

where the γ parameter (which is always non-negative) modulates the influence of the λ parameters on the choice

534 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

536 equal influence on the choice.

SUSTAIN was fit to the SHJ dataset in a supervised fashion, using the same trial order experienced by the

538 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

540 learned rules.