Cerebral Cortex Advance Access published February 26, 2013

Cerebral Cortex
doi:10.1093/cercor/bht014

Quantifying the Internal Structure of Categories Using a Neural Typicality Measure

Tyler Davis[1] and Russell A. Poldrack[1,2,3,4]

[1]Imaging Research Center, [2]Deparment of Psychology, [3]Center for Learning and Memory and [4]Section of Neurobiology, The University of Texas at Austin, Austin, TX 78712, USA

Address correspondence to Dr Tyler Davis, Department of Psychology, University of Texas at Austin, 1 University Station, A8000, Austin, TX 78712, USA. Email: tthdavis@mail.utexas.edu

How categories are represented continues to be hotly debated across neuroscience and psychology. One topic that is central to cognitive research on category representation but underexplored in neurobiological research concerns the internal structure of categories. Internal structure refers to how the natural variability between-category members is coded so that we are able to determine which members are more typical or better examples of their category. Psychological categorization models offer tools for predicting internal structure and suggest that perceptions of typicality arise from similarities between the representations of category members in a psychological space. Inspired by these models, we develop a neural typicality measure that allows us to measure which category members elicit patterns of activation that are similar to other members of their category and are thus more central in a neural space. Using an artificial categorization task, we test how psychological and physical typicality contribute to neural typicality, and find that neural typicality in occipital and temporal regions is significantly correlated with subjects' perceptions of typicality. The results reveal a convergence between psychological and neural category representations and suggest that our neural typicality measure is a useful tool for connecting psychological and neural measures of internal category structure.

Keywords: categorization, fMRI, representation, similarity

Introduction

Categorization is a fundamental process that underlies many aspects of cognition. By grouping alike objects together, categories enable people to generalize knowledge of previously encountered category members to novel ones. Research on categorization seeks to understand how categories are represented and what the implications of this representational structure are for behavior and thought.

Neurobiological research on visual categorization has identified a number of regions that are involved in representing specific aspects of categories. Regions of the visual stream represent information about the features associated with category members (Riesenhuber and Poggio 1999; Palmeri and Gauthier 2004). The prefrontal cortex (PFC) is thought to represent behaviorally relevant aspects of categories such as rules associated with category membership (Ashby et al. 1998; Smith et al. 1998; Miller et al. 2002; Freedman et al. 2003; Ashby and Maddox 2005). Motor and premotor regions may represent habitual responses associated with specific categories (Seger and Miller 2010). The medial temporal lobe (MTL) and subregions of the striatum are thought to bind together aspects of category representations from these other systems. Cortico-striatal visual and motor loops are thought to connect visual stimulus representations with category responses using associative learning (Maddox and Ashby 2004; Seger and Cincotta 2006; Seger 2008; Davis et al. 2012a), whereas the MTLs are thought to represent categories with clusters that join visual information with category-level and behavioral information (e.g., membership, rules) into flexible conjunctions that can be applied across tasks (Love and Gureckis 2007; Davis et al. 2012a, 2012b).

Each of the primary anatomical regions involved in categorization may contain information that can be used to discriminate between categories, a hallmark of category representation that has been emphasized in many recent multivariate and machine learning studies of neural representation (Norman et al. 2006; Diana et al. 2008; Kriegeskorte, Mur, Ruff et al. 2008; Liang et al. 2013). For example, patterns of activation elicited for different objects in regions of the visual stream (Haxby et al. 2001; Spiridon and Kanwisher 2002; O'Toole et al. 2005; Kriegeskorte, Mur, Ruff et al. 2008) and MTL (Diana et al. 2008; Liang et al. 2013) have been found to contain information that can be used to reliably discriminate between many real world object categories. Similarly, neurons in the PFC can discriminate between categories (Freedman et al. 2003) and are theorized to be particularly sensitive to behaviorally relevant differences between stimuli (Pan and Sakagami 2012).

Although coding for differences between categories is one key aspect of category representation, organisms also use categories for a number of other functions such as predicting and inferring features of unseen or novel category members or deciding how characteristic an object is of its category (Markman and Ross 2003). In this way, category representations must not only contain information about how objects differ between categories, but also how objects differ within categories. The manner in which people represent the variability between members within a category is referred to as a category's internal structure (Rosch 1973; Rosch and Mervis 1975). Internal structure allows us to answer such questions as how likely or typical feature combinations (e.g., size, wingspan, and mating habits) are for a given category (e.g., birds), and influences how rapidly and accurately objects are classified (Posner and Keele 1968; Rosch et al. 1976).

Formal cognitive models can offer insight into how internal structure is represented psychologically and in the brain. In many similarity-based categorization models, category members are represented as points in a multidimensional psychological space (Nosofsky 1986; Ashby and Maddox 1993; Kruschke 1992; Love et al. 2004; Minda and Smith 2002; see also Edelman 1998; Gärdenfors 2004). Depending on the specific model, a category representation may be a set of points associated with a given category (exemplar models; Nosofsky 1986), a summary statistic (prototype models; Minda and Smith 2002), or a set of statistics (clustering models; Love et al. 2004; Vanpaemel and Storms 2008) computed over

points associated with a category. A category's internal structure is thought to be reflected in the similarity relations between items and category representations. Category members that are more similar to other members of their category or are nearer to the category prototypes or clusters in a representational space are predicted to be the most typical or likely (Fig. 1*A*).

Formal categorization models are related to mathematical tools that estimate the likelihood or density of particular features or feature combinations from physical or statistical descriptions of objects in the world (Ashby and Alfonso-Reese 1995; Rosseel 2002; Jakel et al. 2009). The key difference between-categorization models and density estimators is that the psychological spaces that underlie categorization are not taken to be veridical representations of the physical world but are rather representations that are affected by psychological factors like learning and attention (Nosofsky 1992; Love 2005). Indeed, although early theories of categorization suggested that psychological internal structure mirrored the physical world, such that objects that are actually the most likely or average for their category are viewed as most typical (Rosch 1973; Rosch and Mervis 1975), there are many

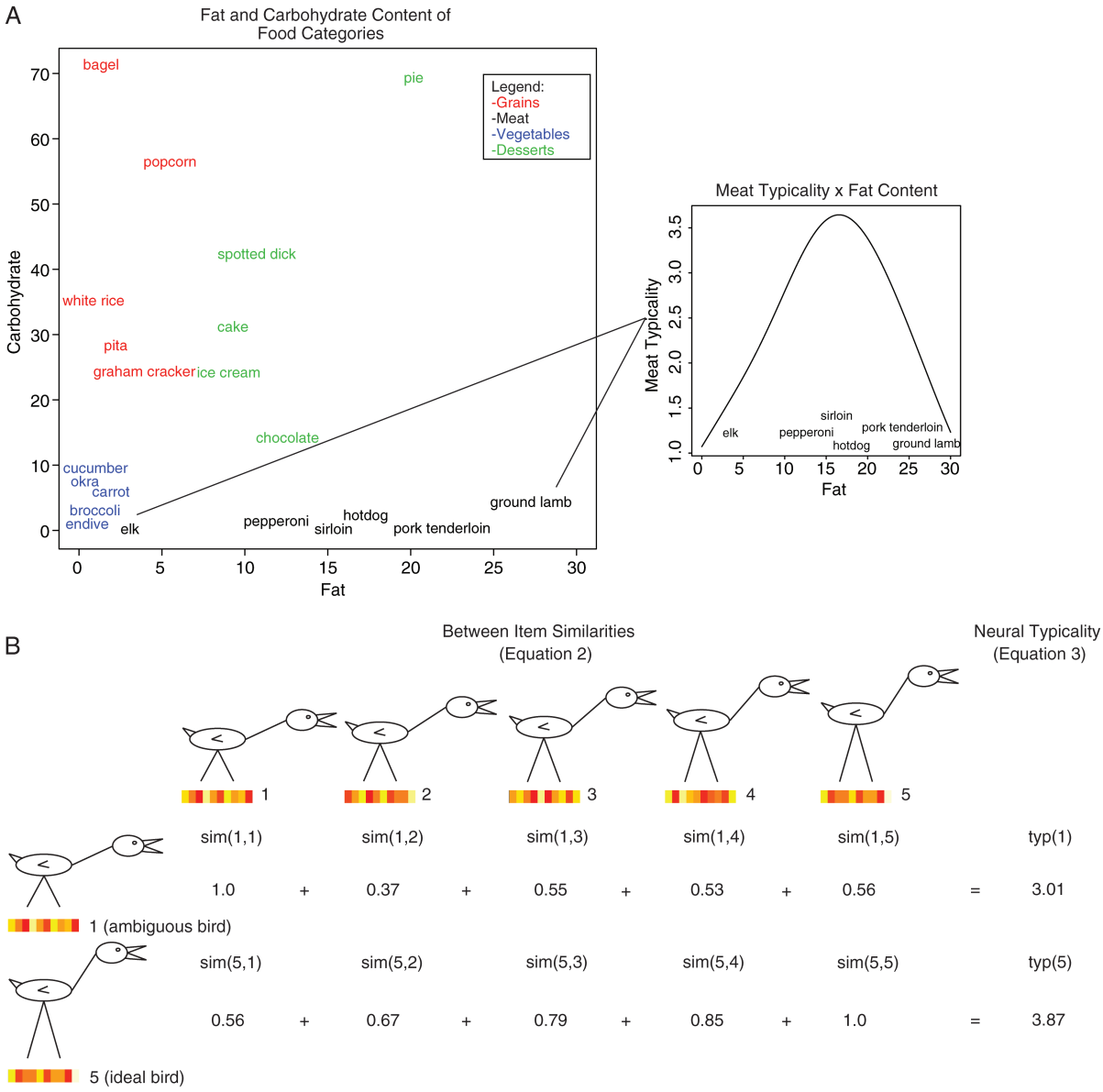

Figure 1. (*A*) (Left) A hypothetical representation of a "Food space" with respect to carbohydrate and fat content dimensions (values are grams per serving). Members of the categories grains, meats, vegetables, and desserts are represented as points in the space. (Right) Gives the typicality (or density) of each meat item, collapsing over the carbohydrate dimension. Typicality estimates for meats were generated from an exemplar model with specificity parameter equal to 1/within meat variance (see Supplementary Material Equations). (*B*) An illustration of the neural typicality measure. The top row of birds is depicted in terms of increasing idealness with respect to category B (ideal is tall and high neck angle). Each bird elicits a pattern of activation given by the red and yellow boxes underneath the bird. In this example, 2 target bird's activation patterns (left), a more ambiguous category member (bird 1), and an idealized category member (bird 5), are compared with each other bird's activation patterns and similarity is calculated. The pairwise similarities are then summed to yield the neural typicality measure. Because bird 5's activation is more similar to other members of the category, it has a higher neural typicality than bird 1.

examples where this relationship does not hold. For example, culture can emphasize particular dimensions of objects differently, leading to different notions of what is typical of a category (Atran 1999; Lynch et al. 2000; Medin and Atran 2004; Burnett et al. 2005). Likewise, contrast (Davis and Love 2010) and goals (Barsalou 1985) can lead people to perceive caricatured or idealized items as typical even though they are not physically average. For example, in the category diet foods, ideal category members are often seen as more typical than physically average members (Barsalou 1985); energy sources (wind, solar, coal, nuclear) can appear as more or less polluting depending on how they are contrasted with one another during learning (Davis and Love 2010). Ongoing research seeks to explain the psychological mechanisms that can transform a physical category space into psychological representations, and thus relate objective physical category structures to psychological ones (Davis and Love 2010). Here, our goal is to develop a method for measuring the internal structure of neural category representations and test how it relates to physical and psychological measures of internal structure.

Categorization models offer insight into how the internal structure of neural representations can be measured in fMRI data. Here, we develop an exemplar-based measure of neural typicality that is based on the similarity between patterns of activation elicited for an object and all other members of its category. This neural typicality measure relates to how exemplar models measure the similarity between an item and other members of its category in psychological space (Nosofsky 1986; Estes 1996) and predict typicality ratings from psychological representations (Nosofsky 1988). Mathematically, this neural typicality measure bears relations to nonparametric kernel density estimators (Ashby and Alfonso-Reese 1995), which are used to predict how likely particular features are given some distribution of objects in the world. A key difference between our measure and related psychological and statistical models is that instead of using psychological or physical exemplar representations, our measure of neural typicality is computed over neural activation patterns, thus giving an estimate of the extent to which an object elicits activation patterns that are like the other members of its category.

Formally, our neural typicality measure is based on the distance d between patterns of activation elicited for a stimulus i and those elicited for other j members of its category (Fig. 1*B*):

$$d_{ij} = \frac{1 - \mathrm{corr}(\beta_i, \beta_j)}{2}, \quad (1)$$

where the β's are trial-by-trial β-series estimates of the pattern of activation elicited for each j stimulus. A Pearson correlation distance metric is used because it normalizes for differences in mean activation level and variability between stimuli. This property makes correlation distance potentially less susceptible to differences in univariate activation between stimuli than other distance metrics (Kriegeskorte, Mur, Bandettini et al. 2008), assuming that the distribution of activation over voxels within an region of interest (ROI) is homogenous and does not reflect a mixture of signals.

The distance between the activation patterns for i and j is transformed to similarity by:

$$S_{ij} = \exp(-d_{ij}), \quad (2)$$

where the exponential instantiates a generalization gradient that determines the form by which similarity decreases as a function of the distance between the patterns of activation for stimulus i and j. Statistically, instituting a generalization gradient reduces the impact of stimuli that are distant from stimulus i on the neural typicality computation. The exponential gradient is a common choice in exemplar models following the finding that the relationship between distance and similarity in psychological spaces tends to be exponential (Shepard's Universal Law; Shepard 1987).

Finally, neural typicality (typ) is computed by summing the pairwise similarities between i and each j member of Category J:

$$\mathrm{typ}(i|J) = \sum_{j \in J} S_{ij}. \quad (3)$$

Objects eliciting patterns of activation that are similar to those elicited by other members of their category are more neurally typical than objects that elicit dissimilar patterns of activation.

By obtaining a measure of an item's neural typicality, or averageness of an item's activation pattern with respect to other members of its category, it is possible to test how different psychological and physical factors influence the internal structure of neural category representations. Here, we test how physical and psychological measures of typicality are reflected in our neural typicality measure in a task in which physical and psychological typicality favor different items. Our primary hypothesis is that neural and psychological representations will be linked such that items judged to be typical by subjects will be those that elicit patterns of activation most like other members of their category.

In the present task, subjects learn, using trial and error, to categorize schematic birds that vary along 2 perceptual dimensions (leg length and neck angle) into 4 categories (Fig. 2*A*). Previous research has found that this type of category-learning task leads to an internal structure that favors physically idealized items (Fig. 2*B*; Davis and Love 2010) because the categories contrast highly with one another. That is, in terms of subjects' psychological similarity space, objects that are more physically idealized or caricatured (e.g., tall birds with a high neck angle in category B; Fig. 2*B*) tend to be viewed as the most typical of their categories and are classified more rapidly and accurately as opposed to those that are actually the most physically similar to other category members. Likewise, when subjects are asked to reconstruct, from memory, an average category member, their reconstructions tend to be caricatured relative to the true category averages (Davis and Love 2010). Together, these results suggest that, in subjects' psychological similarity spaces, idealized items are those that are represented as the most likely or the most similar to other members of their category.

Because the task is highly controlled and its effects on psychological measures of internal structure have been well characterized by previous research (Davis and Love 2010), it is straightforward to make predictions about how internal structure will manifest in the observed neural activation patterns. Specifically, because physical and psychological typicality are dissociable, it is possible to make divergent predictions for how the neural typicality gradients would appear if they reflected physical similarity (i.e., physical averageness; see Fig. 2*C*) versus psychological typicality (Fig. 2*B*).

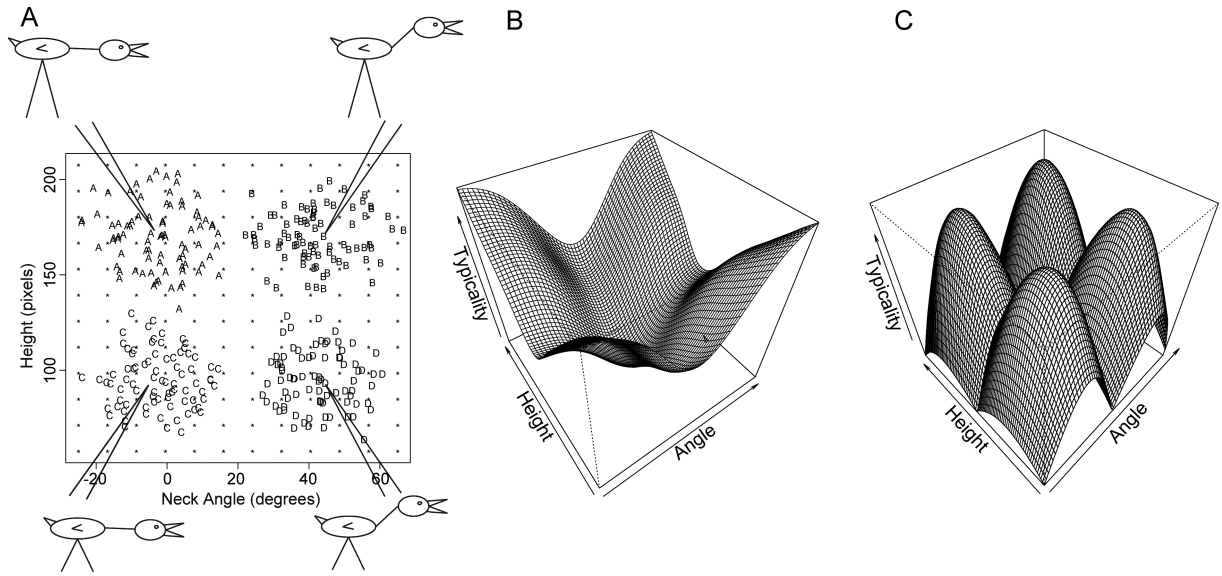

Figure 2. (*A*) Stimuli and category structure for the experiment. Stimuli were birds that varied in terms of their neck angle and height (leg length). During the learning phase, birds were sampled from 4 nonoverlapping category distributions (A, B, C, and D). During the test phase, stimuli were sampled according to a grid that spanned the range of variation observed in the category-learning task (black dots). (*B*) An example of the psychological typicality gradients that are found in the task, plotted in relation to the physical stimulus space. Subjects find objects that are idealized members of their categories typical with respect to their category, and not those that are physically average. (*C*) An example of typicality gradients based on physical similarity. Objects that are average with respect to the physical category spaces are favored because they have the highest similarity to the physical properties of other category members.

Our primary hypothesis is that psychological and neural measures of internal structure will be linked, without regard to where in the brain this might occur. However, given previous literature on object and category representations, it is also possible to make predictions regarding which specific regions will be sensitive to internal structure. Because internal structure is inherently dependent on representing differences between stimuli, the convergence between neural and psychological measures of internal structure is likely to occur in regions of the brain that represent the constituent features of stimuli within a task. In visual categorization, these differences between stimuli are thought be represented in regions of the visual cortex (Freedman et al. 2003; Seger and Miller 2010). Here, we expect that early visual cortex will represent internal structure, because early visual cortex has been theorized to represent simple visual categories in previous categorization research (Reber et al. 1998, 2003; Aizenstein et al. 2000; Zeithamova et al. 2008; for review, see Ashby and Maddox 2005) and is sensitive to primitive features like length and angle (Haynes and Rees 2005; Kamitani and Tong 2005) and information about retinotopic location (Engel et al. 1997). We also expect higher level temporal and medial temporal regions to be sensitive to internal structure because these regions are theorized to bind together features from early visual regions into flexible conjunctive category representations (Davis et al. 2012a, 2012b).

Other regions that are involved in aspects of categorization are thought to be less sensitive to featural differences between stimuli and more to behavioral differences (i.e., differences in responses or rules) or process-level differences (i.e., uncertainty processing) that are not specific to categories or individual stimuli. We employ between-category classification analyses to examine how regions of the brain discriminate between members of different categories and on the basis of behavioral response. Regions of the brain like the PFC and motor/premotor cortex are thought to abstract over differences within categories and be sensitive to behavioral differences between categories (e.g., rules; Freedman et al. 2003; Maddox and Ashby 2004; Pan and Sakagami 2012). We also employ univariate activation measures that are thought to be sensitive to psychological processes (Jimura and Poldrack 2011). One type of process that differs between stimuli within categories and is related to typicality is uncertainty processing. Regions of the ventral striatum are known to be sensitive to entropy or uncertainty processing (Grindband et al. 2006; Davis et al. 2012b) and may correlate with psychological typicality in the present task.

To foreshadow the results, we find that neural typicality significantly correlates with subjects' perceptions of typicality in early visual regions as well as regions of the temporal and medial temporal cortex. These results suggest that neural and psychological representational spaces are linked and validate the neural typicality measure as a useful tool for uncovering the aspects of category representations coded by specific brain regions.

Materials and Methods

Subjects

Seventeen young adult volunteers (ages 18–40 years; 7 females) participated in the study for $25/h compensation. Data from 4 subjects were excluded, 2 due to imaging artifacts and 2 for failing to learn the task. Each participant provided signed informed consent to participate in the study, and all procedures were approved by the IRB of the University of Texas at Austin.

Design

In the experiment, subjects learned how to categorize novel line-drawn birds into 4 categories (A, B, C, and D; Fig. 2*A*). The experiment consisted of a Learning Phase (Fig. 3*A*), a Test Phase (Fig. 3*B*),

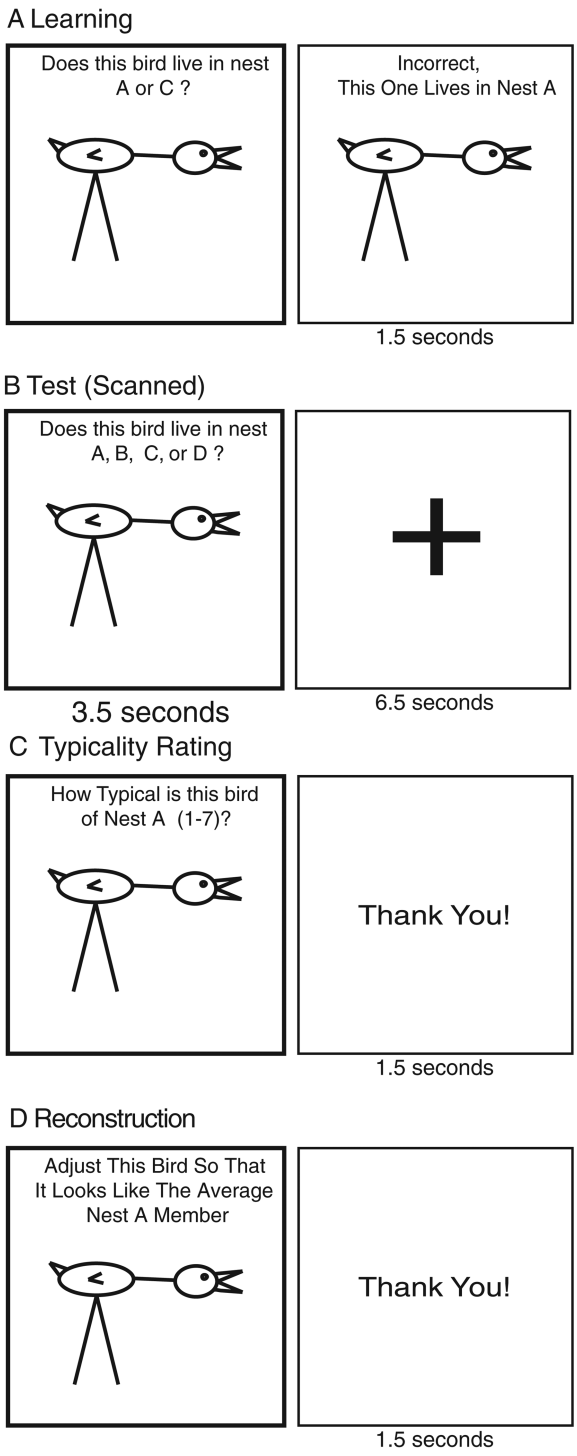

Figure 3. Illustration of each of the task phases. (*A*) Learning phase. On each trial of the learning phase, subjects are presented with a bird drawn from one of the category distributions (blobs of letters in Fig. 3*A*) and are asked to classify it. They receive feedback about their response and the correct category assignment. (*B*) Test Phase. The test phase was functionally scanned. On each trial of the test phase, subjects are presented with a bird drawn from the grid (black dots in Fig. 3*A*) and are asked to classify it. No feedback is delivered in the test phase. Trials are separated by 6.5 s of fixation. (*C*) Typicality rating phase. On each trial of the typicality-rating phase, subjects are presented with a bird drawn from the grid (black dots in Fig. 3*A*) and are asked to rate it in terms of its typicality for a given category. There was no feedback in the typicality phase. (*D*) Reconstruction phase. On each trial of the reconstruction phase, subjects are presented with a bird drawn from the grid (black dots in Fig. 3*A*) and are asked to adjust it (using the arrow keys) so that it matches the average bird in a target category. There was no feedback in the reconstruction phase.

a Typicality-Rating Phase (Fig. 3*C*), and a Reconstruction Phase (Fig. 3*D*). Only the Test Phase was scanned.

Materials

The stimuli were line-drawn birds that differed in terms of their neck angle (degrees from base of neck) and their height (vertical length of legs in pixels from base of body; see Fig. 2*A*). For the category-learning task, the values along the height and angle dimension were sampled for each stimulus from 1 of 4 normal distributions centered at (category: angle, height): (A: −2.5, 170), (B: 32.5, 170), (C: −2.5, 95), (D: 32.5, 95). The standard deviation of the generating distribution for each category was 20° and 33.33 pixels. For each sequence of 36 stimuli (blocks), there were 9 birds from each category, constrained to have a mean equal to that of its generating distribution. Individual stimuli were constrained to be within one standard deviation of their generating category. In the Test Phase, stimuli were sampled from a grid with 12 equally spaced values ranging between −25 and 65° on the angle dimension and between 57.5 and 207.5 on the height dimension, yielding 144 unique stimuli. Stimuli for the Typicality-Rating Phase were a subset (1 of 4) of the stimuli used during the Test Phase.

Behavioral Procedure

Prior to beginning the experiment, subjects were given a thorough description of each of the Learning and Test Phases, including trial timings. In addition, they practiced 1 block of Learning Phase and 4 trials of an example Test Phase outside of the scanner so that they understood the expectations for each task. Subjects were not told that there would be Typicality-Rating and Reconstruction Phases until after scanning.

Learning Phase

The Learning Phase (Fig. 3*A*) was completed during the structural imaging. On each trial of the Learning Phase, a stimulus was presented in the center of the projector screen and subjects were asked to choose the category that it belonged to. To facilitate learning, during the Learning Phase, the number of categories subjects were asked to choose between on a given trial was limited to 2. This manipulation was introduced by Davis and Love (2010; "mixed condition") and is used to speed up the rate of learning while resulting in a graded structure indistinguishable from subjects' who have learned by choosing among all 4 categories on a given trial (Davis and Love 2010; "free condition"). For example, on a trial in which the stimulus was from category A, subjects would only have to choose between A and one of B, C, and D. All pairwise category combinations (e.g., A and C, A and B, and A and D) were queried for each subject. Each category was queried 9 times within a block of 36 stimuli (3 times with each other category). Subjects responded at their own pace using button boxes held in each of their hands. Subjects were instructed to use their right and left middle and index fingers to respond. The mapping of hand/finger to physical category was randomized across subjects. After responding, feedback, including the correct category, was presented for 1.5 s followed by a blank screen for 0.25 s. Subjects completed a minimum of 8 learning blocks and continued training for 16 blocks or until they reached a criterion of 29 of 36 correct within a single block. Subjects failing to exceed 29 of 36 correct within 16 blocks were excluded from further analysis. (The learning procedure allows subjects to focus on only 2 categories per trial and thus has a chance rate of 1/2 for fully random guessing. However, subjects can achieve up to 2/3 correct by focusing on only a single dimension. (e.g., learning A and B are tall and C and D are short). Twenty-nine of 36 represents the upper 95% cutoff of the binomial distribution with a 2/3 probability correct over 36 trials. Thus, subjects achieving 29 of 36 or greater correct within a block of 36 are responding better than would be expected if they had only learned a single dimension and guessed randomly otherwise.)

Test Phase

The Test Phase (Fig. 2*B*) was conducted during functional scanning. On each trial of the Test Phase, a stimulus was presented in the center of the screen and subjects were asked to categorize it into 1 of the 4 categories. All categories were available to choose from during every Test Phase trial. The stimulus would remain on the screen for 3.5 s, during which subjects could respond at any time. After 3.5 s, a fixation cross was presented for 6.5 s. The Test Phase was split into 4 blocks of 36 trials. Stimuli were sampled from the grid in an interleaved sequence (e.g., every fourth stimulus within each category/quadrant starting with the first, second, third, or fourth) to ensure an even sampling of the stimulus space within each block (Supplementary Fig. 1). Trial order within a block was randomized. Each block of the Test Phase was scanned in a separate functional run.

Typicality-Rating Phase

The Typicality-Rating Phase (Fig. 3*C*) was completed after the test phase, outside of the scanner. Prior to beginning the Typicality-Rating Phase, subjects were given instructions to base their typicality ratings on a stimulus' relationship to other members of its category and were given the examples robin and penguin as instances of typical and atypical birds, respectively. On each trial of the Typicality-Rating Phase, a stimulus was presented in the center of the screen and subjects were asked to rate it in terms of how typical it was for its category. For example, for an A stimulus, subjects would be asked, "How typical is this bird of category A (1–7)?" The Typicality-Rating Phase was self-paced. After responding on each trial, subjects were presented with the words "Thank You" for 1.5 s. The Typicality-Rating Phase consisted of 36 trials with the same stimuli as were presented during the final block of the Test Phase. Although only a subset of stimuli was rated in the Typicality-Rating Phase, the interleaved sampling scheme used for sampling from the stimulus grid insured that an even coverage of the category space was achieved for each subject.

Reconstruction Phase

The Reconstruction Phase (Fig. 3*D*) was completed immediately after the Typicality-Rating Phase outside of the scanner. On each trial of the reconstruction phase, subjects used the arrow keys to adjust the neck angle and line-length of example birds to match what they remembered to be the average member of each category. For example, on a category A trial of the reconstruction phase, subjects were presented with a bird and asked to, "Adjust this bird so that it looks like the average category A member." The reconstruction phase consisted of 3 blocks in which each category was queried one time in a random order. The example birds that subjects were instructed to adjust were drawn randomly from their category's quadrant of the test grid. For example, on an A trial, an example bird would be drawn randomly from the quadrant containing category A members. All reconstruction trials were self-paced. After responding on each trial, subjects were presented with the words "Thank You" for 1.5 s.

fMRI Image Acquisition

Data were collected at the Imaging Research Center at the University of Texas at Austin on a 3T GE Signa MRI. Functional images were acquired using a single-shot T_2*-weighted EPI pulse sequence using an oblique axial slice prescription with the following parameters: slice thickness = 3 mm, slices = 32; TR = 2.5 s; TE = 30.5 ms; 64 × 64 matrix; FOV = 220 mm. The first 2 volumes were discarded from each functional time series to allow for T_1 stabilization. To assist in registration, a high-resolution fast spin echo T_2 weighted anatomical image (TR = 3.5 s; TE = 7.9 ms; 256 × 256 matrix; FOV = 280 mm) was collected with the same slice prescription as the functional images. In addition, 2 high-resolution SPGR T1 structural images were acquired in the sagittal plane (slice thickness 1.3 mm; slices = 256; TR = 6 ms; TE = 1.2 ms; 256 × 192 matrix; FOV = 280 mm). Owing to a shim failure that caused image distortions (shearing) in the first functional run/Test Phase for all but 1 subject, the first functional series was discarded in all subjects, leaving 3 functional runs for imaging analysis.

Statistical Methods

Behavioral Analysis and Construction of Internal Structure Measures

For analysis of behavioral responses, response time, and typicality ratings, a distance-to-the-bound variable was constructed that gave each stimulus' overall distance from the boundaries that separate the categories in the stimulus space. Distance-to-the-bound is a useful measure of idealization: items that are distant from the bound are more idealized than items close to the bound (Davis and Love 2010). Because preliminary testing revealed that distance-to-the-bound along both dimensions (angle and height) significantly predicted each of the behavioral measures and the dimensions did not interact, for presentation purposes, both dimensions were combined into an aggregate distance-to-the-bound measure based on a stimulus' additive distance-to-the-bound along each dimension and collapsed into 5 evenly spaced distances. This aggregation did not affect the nature or significance of any statistical results (all distance-to-the-bound tests where significant for both stimulus dimensions). For statistical tests, distance-to-the-bound was regressed on each of the behavioral variables using hierarchical linear models that included random intercepts and distance-to-the-bound slopes for each subject. In the figures, the effect of distance-to-the-bound is depicted with respect to the category B (high angle; high height). Error bars in all figures depict between subject standard errors of the mean for each level of distance-to-the-bound.

Construction of Internal Structure Measures

In order to test how neural typicality related to psychological and physical typicality, we created internal structure measures that reflected these disparate predictions. For the psychological typicality measure, a value for each of the Test Phase stimuli was generated by interpolating, on an individual subjects basis, a predicted typicality rating from the subjects' observed typicality ratings (Fig. 2*B* for a group-level within-category interpolation). Interpolation was accomplished by regressing the observed typicality ratings on each stimulus' physical coordinates using a generalized additive model (GAM; Hastie and Tibshirani 1990; Wood 2006). GAMs are a type of nonparametric regression that model linear or nonlinear relationships between 2 or more variables using a set of smoothing functions and are especially useful for interpolation when, as in the present case, we want to capture subject-level idiosyncrasies in the perception of typicality across the stimulus space while making few assumptions about their form (linear or nonlinear). Using a smoothing function for interpolation also has the benefit of reducing trial-by-trial noise in the psychological typicality measure that would arise from using raw typicality ratings. Although these interpolated psychological typicality measures were based on subsets of the full test phase grid (see Typicality-Rating Phase procedure), statistical tests using hierarchical linear models revealed that increases in interpolated typicality predicted decreases in response time [$t_{(12)} = 5.61$, $P < 0.001$] and increases in probability of choosing the most likely category ($z = 5.08$, $P < 0.001$), 2 other measures of internal structure (Rosch et al. 1976; Davis and Love 2010; see Results for further tests of these measures). A test of the effect of distance-to-the-bound on interpolated typicality ratings revealed that distance-to-the-bound along both dimensions

contributed significantly to the interpolated ratings [height: $t_{(12)} = 3.38$, $P < 0.01$; angle: $t_{(12)} = 2.45$, $P < 0.05$].

For the physical typicality measure (Fig. 2*C*), which predicts that objects will be more typical to the extent that they are physically similar to other members of their category, we used an exemplar-based measure of each item's physical typicality to members of its category presented during the training phase (for fitting details and formalism, see Supplementary Material Equations). Because the form of the psychological typicality measure was allowed to vary between subjects, we also allowed the form of the physical typicality gradient to vary between subjects by fitting an exemplar model to each subjects' individual test phase performance. Three parameters were varied between subjects: an attention parameter that controlled the impact of the height dimension on between item similarities (attention to angle is 1-height), a specificity parameter that controls the width or variance of the similarity gradient between stimuli, and a response scaling parameter which scales the effect of similarity on choice, but does not affect the similarity/typicality itself (Ashby and Maddox 1993). None of these parameters are able to change the basic prediction that physically average items are the most typical. Alternate models that were based only on the objective stimulus distributions (i.e., multivariate Gaussians, kernel density estimators, and physical prototypes) were also tested but yielded equivalent (nonsignificant) results and thus are not included for the sake of brevity.

fMRI Analysis and Image Preprocessing

FMRIB's Software Library (FSL) was used for image processing and standard univariate analysis. For preprocessing, functional time series were skull stripped using BET, corrected for motion using MCFLIRT, and high-pass filtered (cutoff = 100 s).

Multivariate fMRI Analysis

A β-series decomposition (Rissman et al. 2004) employing an LS-S procedure (Mumford et al. 2012) was conducted on the preprocessed, unsmoothed functional time series for each of the blocks during the Test Phase to obtain trial-by-trial estimates of hemodynamic response. LS-S iteratively models a β estimate for each stimulus in the task by computing, one at a time, a separate least-squares model for each stimulus, while simultaneously controlling for the effect of the other trial onsets using a single stimulus regressor. Simulations on real and artificial data reveal that this method outperforms many commonly employed β-series extraction methods (e.g., ridge regression) and obtains accurate estimates of the hemodynamic response at even shorter ITIs than used in the present experiment (Mumford et al. 2012). Fluctuations in the duration of the hemodynamic response due to trial-by-trial differences in time-on-task were controlled for in the LS-S model by including a regressor in which the duration of the hemodynamic response varied according to subjects' response time. This response time regressor was orthogonalized with respect to the hemodynamic responses predicted by the fixed duration trial onsets alone. Unconvolved motion parameters were included as nuisance variables. The β-series were registered to the Montreal Neurological Institute (MNI)-152 template using the registration parameters from the univariate FEAT analysis (described below).

Neural Typicality Analysis

The computed β-series were used as inputs into the neural typicality measure defined in the Introduction. The spatial localization of the β-series used to compute the neural typicality measure was selected using a search light algorithm (Kriegeskorte et al. 2006) with a 3-voxel radius. The neural typicality values for each stimulus, computed over the voxels within each searchlight, were then correlated with the different predictions for internal structure (psychological or physical typicality measures described above) using a Pearson correlation. The obtained subject-level correlation maps were transformed using Fisher's Z transformation and combined for second-level between-subjects analysis. Subject-level maps were submitted to a group permutation test using FSL's "randomize" function (10 000 repetitions) with spatial 5-mm FWHM variance smoothing, and corrected for multiple comparisons using a cluster forming threshold of $t > 2.18$ ($P < 0.05$) and corrected extent threshold of $P < 0.05$. For consistency with the behavioral measures, we plotted the observed neural typicality in significant clusters (taken from 6-mm spheres around Subjects' peaks of activation) as a function of distance-to-the-bound. Although these plots and associated statistics (not presented) are biased due to having been selected on the basis of whole-brain statistics (Kriegeskorte et al. 2009; Vul et al. 2009), they nonetheless help to illustrate consistency across behavioral and neural measures, the magnitude of the effect, and the linearity assumption of the Pearson correlation is valid.

Partial Activation Analysis

We also conducted additional analyses partialling out the effect of activation and physical or psychological typicality from each searchlight prior to correlating the respective internal structure measure with neural typicality. This mean signal correction has been argued to neutralize differences in pattern similarity between stimuli that scale with activation level or variance, but are not entirely corrected for by correlation's normalizing for mean and variance (Xue et al. 2012). Specifically, for the "partial activation analysis," we partialled out the effect of the mean activation across each searchlight from the neural typicality gradient prior to correlating it with the psychological or physical typicality measures. Likewise, for the "partial physical analysis," we partialled out the effect of physical typicality from neural typicality prior to correlating it with psychological typicality (and vice versa).

Similarity to Opposing Category Analysis

We present a targeted (within the ROIs identified in whole-brain neural typicality analysis) between-category analysis that is a direct extension of our neural typicality measure, but is based on a stimulus' similarity to all (3) other categories that it is not a member of. Aggregating the similarity to opposing category measure over unambiguous category members (Fig. 8*A*) and comparing it to within-category similarity (i.e., Neural Typicality), we also test whether unambiguous items of each category are more similar to members of their own category than members of opposing categories.

Multidimensional Scaling Analysis

Multidimensional scaling (MDS) is a measure of what latent dimensions account for variance in a dissimilarity matrix. We test whether regions identified in a univariate task versus baseline comparison (The MDS analysis is restricted to

regions identified in independent univariate analysis instead of our ROIs identified in the Neural Typicality Analysis because within-category similarity can impact MDS results.) (see below; Supplementary Material) code the physical stimulus dimensions that separate the categories via MDS of the pairwise correlation distance between stimuli (see Kriegeskorte, Mur, Ruff et al. 2008; Kriegeskorte, Mur, Bandettini et al. 2008). For the MDS analysis, we restricted the analysis to stimuli within the test grid that were unambiguous members of their category (stimuli that are distant from the category boundaries) so as to attempt to remove the majority of variance associated with within-category differences between stimuli (e.g., neural typicality) from the MDS results.

Classification Analysis

An epsilon-insensitive linear ν-support vector machine (SVM) was trained to predict subjects' behavioral responses during the test phase using the β-series representations of stimuli within the task. A nested leave-one-run-out cross-validation was employed in which we iteratively trained SVMs on activation patterns from part of the data (all but 1 run) and then tested the trained SVM's ability to predict subjects' responses in the left-out run. Within each training set, an additional (nested) leave-one-run-out cross-validation was used to select the value for the SVM's hyperparameter (ν) that minimized transfer error. Ten equally spaced values of ν on the range (0.1, 1) were tested within each nested cross-validation. The SVM with the best fitting ν parameter was then retrained on the full training set and used to predict subjects' categorization responses in the left out test set. Because SVMs can be biased if the number of training examples differs across classes, all categories were constrained to have equal numbers of responses within a scanning run. The full nested cross-validation procedure was used to calculate mean cross-validated accuracy for each voxel using a searchlight (radius = 3 voxels). Chance performance (0.25) was subtracted from each voxel within the subject-level accuracy maps. For the sake of computational tractability, the SVM analysis was done in native space, and the resulting subject-level accuracy maps were registered to the MNI template using the registration parameters from the univariate FEAT analysis. The subject-level normalized accuracy maps were submitted to second-level group analysis using FSL's randomize with the same thresholding as for the neural typicality analysis. The resulting group-level maps give clusters in which the mean cross-validated accuracy is significantly greater than chance between subjects.

Univariate fMRI Analysis

Voxelwise univariate analysis was conducted using a standard 3-level analysis in FEAT. In the first level, functional data were prewhitened using FILM, high-pass filtered (cutoff = 100 s), and spatially smoothed with a 5 mm FWHM Gaussian kernel. Statistical analyses were performed under the assumptions of the general linear model (GLM). Regressors included the trial onsets, convolved with a double gamma hemodynamic response, and its temporal derivatives. Unconvolved motion parameters and their temporal derivatives were included as nuisance predictors. Response time was controlled for in the same manner as described for the multivariate analysis.

Three separate univariate analyses were conducted to assess which regions were 1) significantly correlated with psychological typicality (using the psychological internal structure measure), 2) significantly correlated with physical typicality (using subject-specific physical typicality measure), and 3) significantly activated/deactivated relative to baseline (fixation; see Supplementary Material Results). First-level statistical maps were registered to the MNI-152 template using 7 DOF to align the functional image to the structural image, and 12 DOF to align the structural image to the MNI-152. Second-level fixed effects analysis combined Test Phase runs within a single subject. Third-level random effects modeling, using Feat's FLAME 1, combined second-level results across subjects. Results were thresholded using a cluster-forming threshold of $Z > 1.96$ and corrected extent threshold using Gaussian random field theory ($P < 0.05$).

Results

Behavioral Results

Learning Phase

All subjects reached the learning criterion of 81% correct within the first 8 blocks of learning. The mean classification accuracy increased across learning blocks starting at 0.78 correct in block 1–0.93 correct in block 8, $t_{(12)} = 5.23$, $P < 0.001$.

Test Phase

The proportion of trials in which subjects chose the most likely category (assuming perfect knowledge of boundaries) was constant across the 4 test runs (0.78, 0.77, 0.78, and 0.79). Whether or not subjects chose the most likely category, however, differed depending on the stimulus' distance to the boundaries that divide the categories in the stimulus space (e.g., the central axes that divide A from C and A from B; Fig. 2*A*). A hierarchical linear model revealed that subjects' probability of choosing the most likely category increased 0.10 for each unit increase in distance-to-the-bound (i.e., as stimuli became more idealized), $t_{(12)} = 9.44$, $P < 0.001$ (Fig. 4*A*). Similarly, response times decreased by 163 ms for each unit increase in distance-to-the-bound, $t_{(12)} = 7.94$, $P < 0.001$ (Fig. 4*B*).

In an additional analysis, we examined subjects' confusion matrices (which categories subjects chose when they did not choose the most likely category) and found that confusions were more likely to occur between categories separated along a single dimension (e.g., A and B or A and C) then for categories separated along both dimensions (e.g., A and D; see Fig. 2*A*). As a proportion of total confusions, 44% of confusions were between categories separated along the height dimension, 50% of confusions were between categories separated along the angle dimension, and 6% of confusions were between categories separated along both dimensions. These results are consistent with the distance-to-the-bound results (categories separated by 2 dimensions are more distant) and suggest that both dimensions contribute independently to subjects' choice behavior.

Typicality Results

A hierarchical linear model revealed that subjects' typicality ratings increased by 0.35 typicality units (1–7) for each unit increase in distance-to-the-bound, $t_{(12)} = 4.61$, $P < 0.001$ (Fig. 4*C*). These results suggest that, as predicted, the stimuli

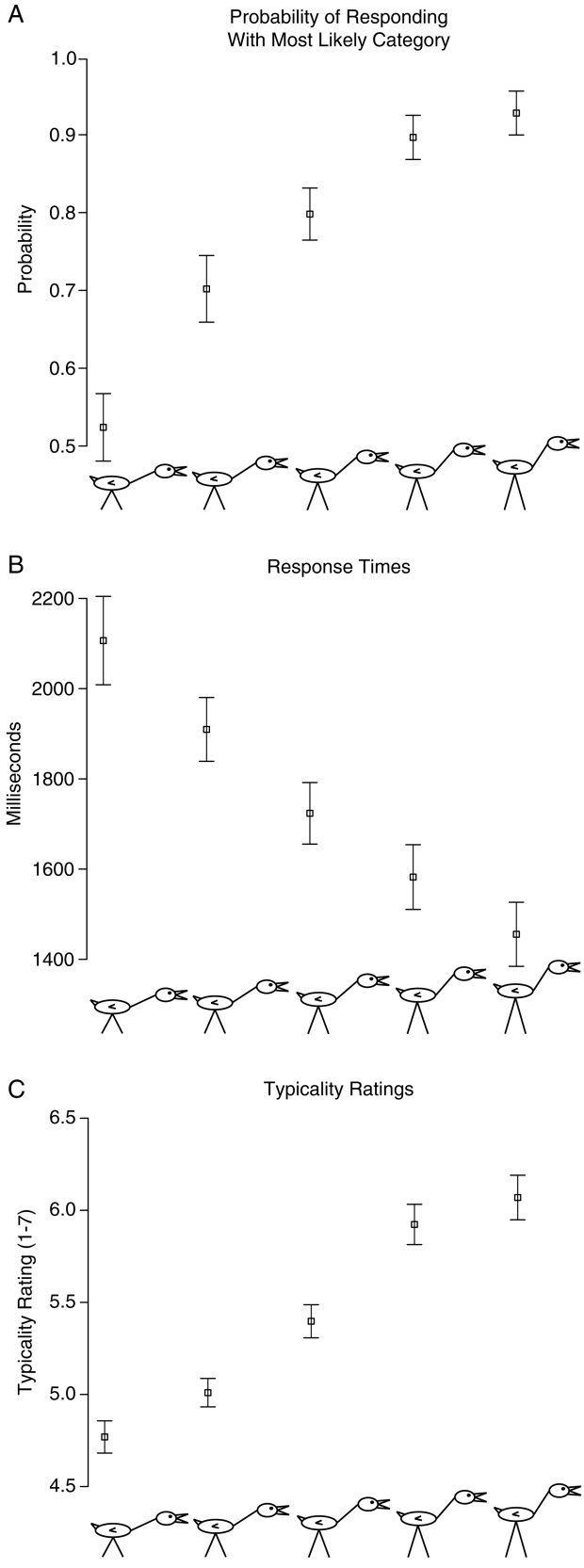

Figure 4. Behavioral results for (*A*) probability of choosing the most likely category, (*B*) response times, and (*C*) typicality ratings as a function of distance-to-the-bound or increasing idealness (from left to right). For presentation purposes, distance-to-the-bound is depicted with respect to increasing idealness in category B (ideal is tall and high neck angle). Error bars represent ± standard error of the mean.

that are perceived as most typical are idealized or caricatures and not the stimuli that were physically average.

Reconstruction Results

For analysis of the reconstruction data, subjects' reconstructions were centered around the true category averages such that a value of 0 would indicate an exact reconstruction of the category average and a positive value would indicate an idealized reconstruction. Every subject's reconstructed category averages were positive, or idealized, along both dimensions. For the angle dimension, subject's reconstructed stimuli were, on average, 0.64 standard deviations greater than the true category averages, $t_{(12)} = 13.76$, $P < 0.001$. For the height dimension, subject's reconstructed stimuli were, on average, 0.75 standard deviations greater than the true category averages, $t_{(12)} = 7.48$, $P < 0.001$. As with the other behavioral measures, the reconstruction results suggest that the physical averages are not average with respect to subjects' psychological representations.

Imaging Results

Neural Typicality is Linked to Psychological Typicality

Our primary focus was on how the internal structure of the categories would be reflected in our neural typicality measure. Our neural typicality measure is based on similarities between multivariate patterns of activation elicited for stimuli in the task. Stimuli that elicit activation patterns that are like other members of their category are more neurally typical than those that elicit dissimilar patterns of activation. One hypothesis was that the internal structure of the neural category representations would be linked to their psychological structure such that the stimuli that are most neurally typical would be those that subjects find most typical. According to this hypothesis, the stimuli that are idealized or caricatured relative to the physical category space should be those that elicit patterns of activation that are most similar to other category members.

A searchlight algorithm was employed to examine how pattern similarity related to subjects' typicality ratings in different regions of the brain. In each searchlight, we computed the multivariate neural typicality measure for each stimulus and correlated the resulting neural typicality gradient with subjects' typicality ratings (Fig. 1*B*). Two clusters exhibited the predicted correlation between neural typicality and psychological typicality: as perceived stimulus typicality increased, the similarity between patterns of activation for the stimulus and other members of its category increased (Fig. 5*A*; Table 1). One cluster spanned the early visual cortex from the lingual gyrus and occipital pole to the superior lateral occipital cortex (cluster-corrected $P = 0.006$; Fig. 6*A*). These early visual regions have been hypothesized to code prototype-based representations of simple perceptual categories in previous fMRI research (Reber et al. 1998, 2003; Aizenstein et al. 2000; Zeithamova et al. 2008; for review see Ashby and Maddox 2005), and are known to be sensitive to primitive object features like the neck angle and leg length dimensions that constitute our stimulus set (e.g., Haynes and Rees 2005; Kamitani and Tong 2005). A second cluster spanned the right posterior MTL regions in the parahippocampal cortex and neighboring hippocampus and posterior into temporal fusiform and lingual gyrus (cluster-corrected

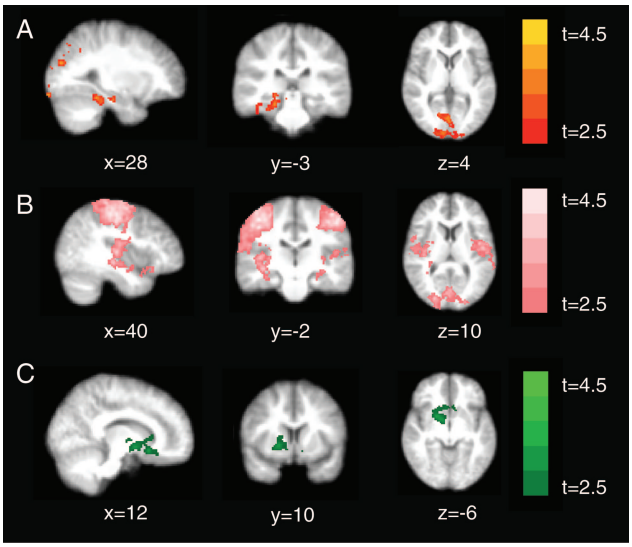

Figure 5. Neuroimaging results for (*A*) regions in which neural typicality measure correlates with psychological typicality(red/yellow); (*B*) regions in which the SVM's cross-validated classification accuracy was significantly greater than chance (pink), (*C*) regions in which univariate activation correlates significantly with psychological typicality (green).

Table 1
Observed clusters for each of the main analyses

Analysis	Regions	Cluster size	*x*	*y*	*z*	Peak (*t*)
Neural typicality measure (psychological typicality)	Early visual cortex	2234	24	−98	−16	4.97
	Right temporal fusiform, medial temporal lobes, lingual gyrus	993	24	−36	−16	4.18
SVM predicting subjects' responses	Right motor, premotor, insula, frontal, and hippocampus	7976	44	−20	54	4.88
	Early visual cortex, occipital fusiform, and lingual gyrus	5660	20	−68	−18	4.22
	Left insula, frontal and hippocampus	3871	−8	48	−20	4.44
	Left motor and premotor cortex	2888	−50	−22	50	4.63
Univariate activation (psychological typicality)	Striatum	866	18	8	−2	3.26

Coordinates (mm) listed are for cluster peak.

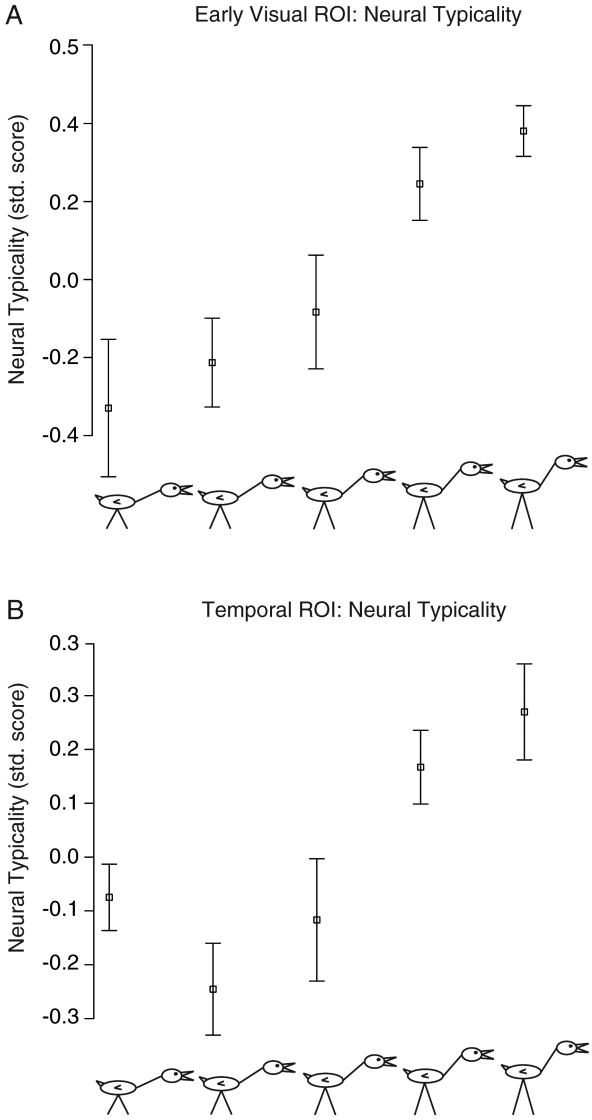

Figure 6. Neural typicality within each whole-brain ROI presented as a function of distance-to-the-bound or increasing idealness (from left to right). For presentation purposes, distance-to-the-bound is depicted with respect to increasing idealness in category B (ideal is tall and high neck angle). Neural typicality scores are standardized within subjects to neutralize individual differences in scale. Error bars represent ± standard error of the mean.

$P = 0.047$ (In an earlier version of this analysis using fewer [5000] iterations in our permutation tests, this MTL cluster's P value was slightly higher [cluster-corrected $P = 0.051$].); Fig. 6*B*). These MTL regions have recently been associated with model-based measures for retrieval of category representations from memory (Davis et al. 2012a, 2012b). The temporal fusiform is often associated with representing higher order features and feature combinations that are constituents of many natural object categories (Martin et al. 1996; Haxby et al. 2001; Palmeri and Gauthier 2004). These results suggest that, in the present task, the internal structure of neural category representations in temporal and occipital regions are linked to subjects' psychological category representations such that objects that are idealized or physical caricatures of their category elicit patterns of activation that are most (mathematically) similar to other members of their category.

Relationship Between Neural Typicality and Psychological Typicality is Not Related to Univariate Activation Differences Between Stimuli

Although the neural typicality measure normalizes with respect to overall activation within a searchlight, and thus mean activation level does not figure directly into any of the similarity calculations, previous research has suggested that it is also important to assess whether pattern similarities are statistically independent of results predicted by mean activation-level (Xue et al. 2010, 2012). Partialling out mean activation can aid in interpreting how univariate activation and pattern similarity results differ and potentially neutralize any residual impact of univariate activation that is not accounted for by using a correlation distance metric. To this end, we also conducted a set of analyses in which we partialled activation out of the neural typicality measure prior to

correlating it with the psychological typicality ratings, yielding statistical maps in which the effect of psychological typicality on neural typicality is statistically independent of mean activation. Consistent with the prediction that the neural typicality measure reflected representational differences between stimuli, both clusters reached statistical significance in the partial activation analysis: early visual ROI $P = 0.013$; Temporal ROI $P = 0.045$ (Supplementary Fig. 2*A*).

Neural Typicality Gradient Does Not Derive From Physical Similarity

Another potential influence on the internal structure of neural category representations is physical typicality. To test the hypothesis of whether physically typical stimuli would be the most neutrally typical and elicit activation patterns that are most similar to other members of their category, we constructed a physical typicality measure that favored physically average stimuli (Fig. 2*C*). In contrast to our pattern similarity analysis that used subjects' typicality ratings and favored ideal stimuli, the physical typicality measures did not reveal any correlation with the neural typicality measure that passed corrected thresholds (most significant cluster: $P = 0.328$). These results suggest that, in the present task, physical similarity is not a significant contributor to the internal structure of neural category representations, at least not at a level that is amenable to detection using fMRI.

Partial Correlation Analysis of Neural Typicality

Although our presentation of internal structure and how it is manifest in neural similarities makes physical and psychological typicality appear as if they are separate hypotheses for how the brain organizes category representations, in many cases, it might be useful to view internal structure from a multiple regression standpoint where these factors (and others) are seen as different influences on how category representations are organized. This is the approach taken by Barsalou (1985), who was one of the first to demonstrate an effect of ideals on internal structure. In this spirit, we conducted additional partial correlation analyses examining whether the effect of psychological typicality was significant after partialling out the effect of physical typicality and vice versa.

The results were highly consistent with the analyses examining the internal structure measures independently. The psychological internal structure measure correlated significantly with neural typicality in clusters in the early visual ($P = 0.007$) and temporal cortices ($P = 0.017$) after partialling out the effect of physical typicality. Physical typicality remained a nonsignificant predictor of neural typicality after partialling out the effect of psychological typicality (most significant cluster: $P = 0.134$).

Similarity to Opposing Categories

One important question in relation to category representation is the extent to which stimuli are represented as distinct from members of other categories and whether this differs as a function of an item's physical or psychological typicality. With respect to between-category similarity, both physical and psychological typicality measures predict that idealized items will be less similar to opposing categories in terms of the present stimulus space. Assuming that the regions that code internal structure also represent between-category differences, dissimilarity to opposing categories as a function of distance-to-the-bound should be evident in the early visual and temporal ROIs. Accordingly, the results of a hierarchical linear model regressing distance-to-the-bound on between-category similarity revealed that similarity to opposing categories decreased as distance-to-the-bound increased in the early visual ROI, $t_{(12)} = 3.91$, $P = 0.002$ (Fig. 8*A*). The effect of distance-to-the-bound on between-category similarity was marginal in the temporal ROI, $t_{(12)} = 2.00$, $P = 0.069$ (Fig. 7*B*). A post hoc analysis that was inspired by the plots of similarity to opposing categories as a function of distance-to-the-bound revealed that there was also a significant quadratic trend in both regions whereby similarity to opposing categories tended to rise for the most extreme/idealized category members (early visual ROI: $t_{(12)} = 3.82$, $P = 0.002$; temporal ROI: $t_{(12)} = 2.45$, $P = 0.031$). The overall decreasing linear trend remained robust in both ROIs after taking into account the quadratic effect (early visual ROI: $t_{(12)} = 3.41$, $P = 0.005$; temporal ROI: $t_{(12)} = 2.30$, $P = 0.04$).

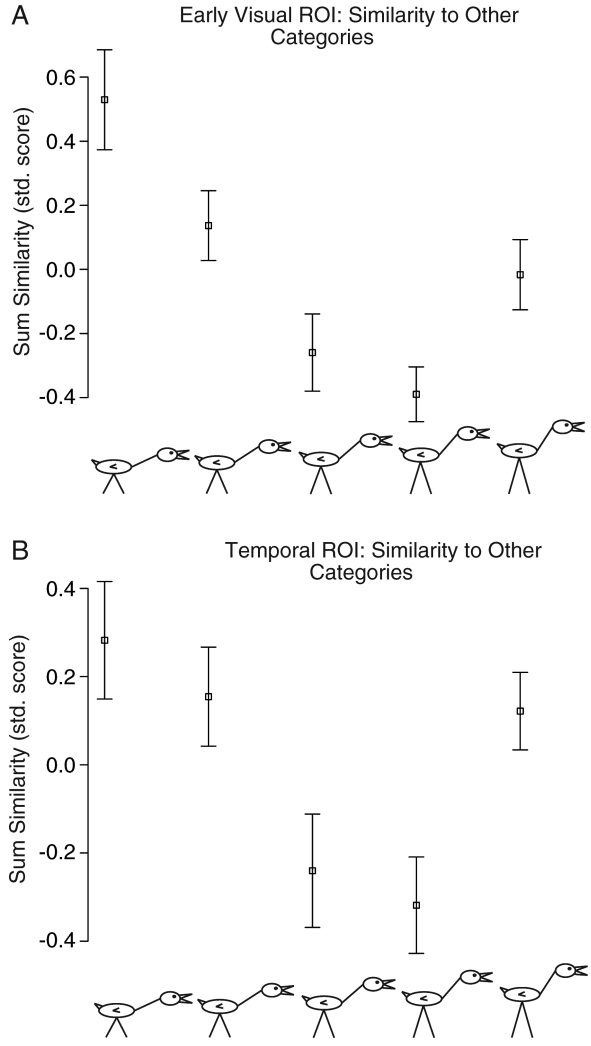

Figure 7. Between-category similarity within each whole-brain ROI presented as a function of distance-to-the-bound or increasing idealness (from left to right). For presentation purposes, distance-to-the-bound is depicted with respect to increasing idealness in category B (ideal is tall and high neck angle). Between-category similarity scores are standardized within subjects to neutralize individual differences in scale. Error bars represent ± standard error of the mean.

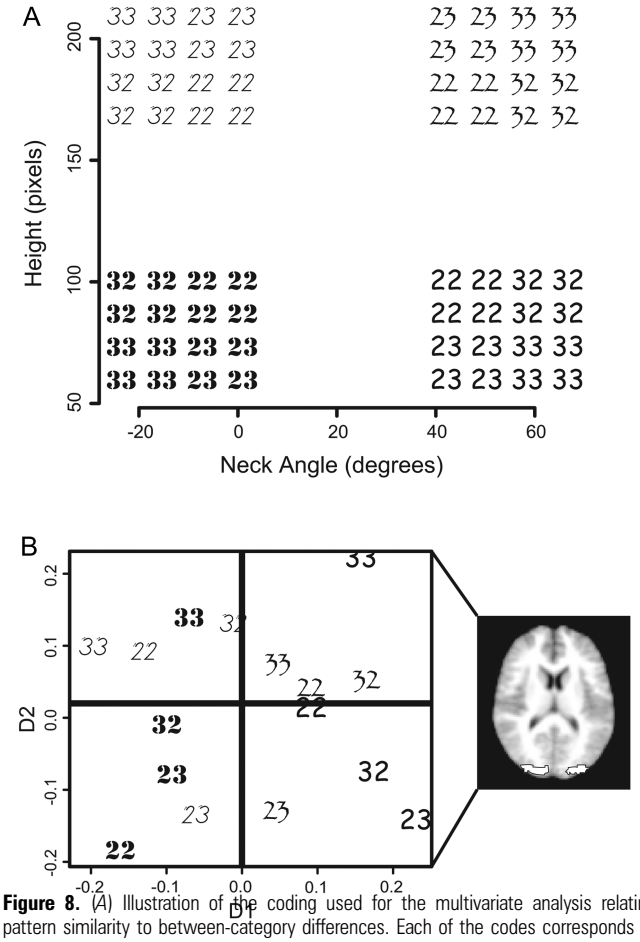

Figure 8. (*A*) Illustration of the coding used for the multivariate analysis relating pattern similarity to between-category differences. Each of the codes corresponds to a group of stimuli presented during the test phase. The codes are structured based on the distance to the boundary that separates the categories (center of the stimulus space). The distance from the boundary along each dimension is divided into 3 equal sized sections (1–3), with 3 being the furthest from opposing categories or most ideal. Only the most unambiguous stimuli (2 and 3 s on both dimensions) were included so as to focus on between-category differences. (*B*) Results from multidimensional scaling of between-category pattern similarity in the occipital pole.

According to models of categorization (Nosofsky 1986; Kruschke 1992; Ashby and Maddox 1993; Minda and Smith, 2002; Love et al. 2004), both similarity to opposing categories and within-category similarity should affect categorization judgments. Typical objects should not only be more like members of their own category and less like members of other categories, but also within-category similarity should be stronger overall than between-category similarity. To test this hypothesis, we directly compared mean within-category similarity (i.e., neural typicality) for unambiguous items to mean between-category similarity for unambiguous items. In both ROIs, within-category similarity was significantly greater than between-category similarity [early visual ROI: $t_{(12)} = 2.35$, $P = 0.037$; temporal ROI: $t_{(12)} = 2.37$, $P = 0.036$].

Multidimensional Scaling

Another method for characterizing whether a brain region codes differences between categories, with respect to many tasks, is to examine whether regions of the brain code the dimensions of variation that separate categories. In the present task, there are 2 independent dimensions of variation for all stimuli (neck angle and height). Although physical differences along these dimensions cannot account for neural typicality/within-category similarities between stimuli (there is no linear reweighting of the physical stimulus dimensions that would make ideal items more average than physically average items), it is still possible that the larger between-category variation along these dimensions is detectable. As an exploratory technique, classical MDS is well suited for discovering independent latent dimensions of variation that explain the (dis)similarities between stimuli in a multidimensional space. To examine the question of whether any of the ROIs identified in the univariate (task vs. baseline contrasts; see Supplementary Material) coded the stimulus dimensions that separate the categories, we grouped the most highly typical and unambiguous stimuli in each category into groups based on their distance from opposing categories (Fig. 8*A*) and computed representational dissimilarity matrices that coded the pattern (dis)similarity between each group. An MDS of the representational dissimilarity matrix extracted from the occipital pole ROI revealed a 2D solution that corresponded well to the physical stimulus dimensions (Fig. 8*B*). Although this 2D solution only accounted for 25% of the variance in the dissimilarity matrix, it nonetheless accurately reflected the physical between-category differences for the different stimulus groupings. Indeed, 75% classification accuracy can be achieved by using decision boundaries orthogonal to the 2 extracted dimensions (formal linear discriminant analysis achieved the same level of accuracy). One weakness of MDS is that, like factor analysis, independent component analysis, and other exploratory data analysis techniques, the dimensions that MDS extracts are subject to interpretation and often may not be interpretable at all (Venables and Ripley 2002). Indeed, no dimensions extracted in other regions produced easily interpretable results or separated the categories significantly above chance in linear discriminant analyses.

Classification Analysis

Some regions of the brain may discriminate between categories on the basis of behavioral factors (e.g., responses and rules; Freedman et al. 2003; Maddox and Ashby 2004; Pan and Sakagami 2012) and thus would not exhibit graded effects (i.e., similarity to opposing categories analysis) or may not separate the categories on the basis of visual stimulus dimensions (MDS). Thus, we also include a whole-brain, between-category classification analysis that attempts to predict subjects' categorization responses using SVMs. Unlike the previous analyses, SVMs are sensitive to between-category information that can be coded in a variety of forms. The SVM's measure of between-category representation is its ability to accurately predict (classify) subjects' categorization responses from the β-series activation patterns elicited for stimuli in the task. To this end, the SVMs were trained on a group of β-series patterns (all but 1 scanning run) and then used to predict subjects' responses during the left out scanning run on the basis of this training. The SVMs are given no information about the underlying stimulus space, and unlike the MDS analysis, do not make any assumptions about how the dimensions that separate the categories will be organized. Thus, the SVMs can be sensitive to regions that code rule-based or behavioral differences between categories, regions that encode information about their perceptual differences, or regions that code some combination of behavioral and perceptual information.

Clusters in widespread regions of the visual cortex and occipital fusiform ($P = 0.042$) and right motor/premotor cortex, insula, hippocampus, and lateral frontal cortex ($P = 0.028$) were found to discriminate between categories significantly greater than chance in the SVM analysis (Fig. 5*B*; Table 1). Two separate clusters covered the same general motor/premotor, insula, hippocampus, and frontal regions in the left hemisphere, but were marginally significant (motor/premotor $P = 0.089$; insula, hippocampus, and frontal $P = 0.065$).

Although there is strong overlap in the visual and MTL regions that discriminate between categories and represent internal structure, the motor/premotor, insula, and frontal regions were only identified in the between-category analysis. These results are consistent with the hypothesis that PFC and motor/premotor regions are more sensitive to behavioral aspects of categories (Freedman et al. 2003; Maddox and Ashby 2004; Pan and Sakagami 2012). However, because behavioral responses are strongly associated with the perceptual characteristics of each category, the SVM results are also consistent with the hypothesis that these regions contain some perceptual information about the categories.

Univariate Typicality Models Predict Activation in Ventral Striatum

As discussed above in the partial activation analysis, univariate activation that is significantly related to a particular measure or contrast is often viewed as an indicator that a region is involved with processes associated with the measure (as opposed to representations per se). Thus, we also examine how our internal structure measures (psychological and physical typicality gradients) relate to univariate activation, as many psychological processes are sensitive to typicality or goodness of membership (Murphy 2004).

In the present context, in relation to the goals of the test phase, psychological typicality or atypicality, which favors idealized stimuli, may reflect subjects' level of certainty in an objects' category membership. Highly typical stimuli are almost certain to be members of their category whereas atypical stimuli are associated with greater uncertainty. Accordingly, while we did not find any regions that positively correlated with typicality, we did observe a cluster of activation centered in the putamen and spanning the surrounding ventral striatum that negatively correlated with typicality (see Fig. 5*C*; Table 1). These results are consistent with previous results suggesting a role for the ventral striatum and dopamine system in processing categorization novelty and uncertainty (Grindband et al. 2006) or entropy (Aron et al. 2004; Davis et al. 2012b). They also highlight the unique utility of multivariate pattern analyses for understanding relations between cognitive and neural representations. We did not find any regions in which univariate activation tracked physical typicality (favoring physically average stimuli).

Discussion

The present study sought to decode the internal structure of category representations using multivariate patterns of activation measured with fMRI. To this end, we coined a neural typicality measure that is based on the similarity between the patterns of activation elicited for a stimulus and other members of its category. According to this measure, objects that elicit patterns of activation that are more similar to other members of their category or are in regions of higher density with respect to a neural category space are more neurally typical. By applying the neural typicality measure to imaging data from an artificial categorization task, we tested whether the neural representations of categories in an artificial category-learning task reflected subjects' psychological perceptions of typicality (which would favor physically "ideal" category members), or were instead organized on the basis of physical typicality (which would favor physically "average" category members). We found, in regions of the temporal and occipital cortex, that psychological and neural typicality were closely aligned, such that neural typicality increased as a function of subjects' perceptions of typicality. In contrast, a physical typicality measure that favored physically average category members was not significantly associated with neural typicality gradients in any brain region. These results suggest that patterns of activation in the brain can carry fine-grained information on the internal structure of subjective category representations, a topic that has been at the center of debate in cognitive research for decades.

The present research adds to the growing consensus that categorization depends on interactions between a number of different brain regions (Ashby and Maddox 2005; Poldrack and Foerde 2008; Seger and Miller 2010). Indeed, as we found in our series of both between- and within-category (and univariate) analyses, information about category structure and membership is widely distributed throughout the brain. An important point that this observation highlights is that there may not be any brain region that can be thought of representing all aspects of categories, and thus it might be most accurate to think of brain regions in terms of the aspects of category representations that they code. In this regard, our neural typicality measure may be a valuable additional tool for studying category representations using neuroimaging. Because internal structure depends, at least in part, on information about featural variation within a category, neural typicality may be more uniquely sensitive to representations in brain regions that process a category's constituent features than are between-category techniques like classifiers. Indeed, between-category information may not only code featural differences between categories, but also a number of other more behavioral differences. Together, multivariate techniques that measure both between and within-category information (in addition to univariate analyses) are able to reveal a broader picture of how different brain regions contribute to categorization.

One advantage of our neural typicality measure over other measures of neural function that have been used to study categorization is that it brings fMRI analysis more in line with theoretical mechanisms posited by cognitive models. Like our measure, cognitive models predict that categorization depends on processes that compute similarity or matches between items and category representations (Nosofsky 1986; Kruschke 1992; Ashby and Maddox 1993; Minda and Smith 2002; Love et al. 2004). Insofar as neural activation patterns elicited in the task are measures of the underlying neural representation, our results provide strong evidence that the brain may engage the same types of processes as posited by cognitive models. Evidence that these neural activation patterns differentiate items both within- and between-categories is consistent with this representational interpretation. Univariate measures of activation are often less interpretable and may

correlate with psychological measures for reasons other than representational similarity or matching processes. For example, in the present context, the deactivation of regions of the striatum with increasing typicality likely indicates an uncertainty signal, as opposed to category representation, based on current knowledge of the ventral striatum's role in categorization (Grindband et al. 2006; Davis et al. 2012b).

One univariate activation-based measure of neural representation that has recently been theorized to be sensitive to internal category structure is repetition suppression. Repetition suppression measures representational similarity as the degree of univariate neural adaptation between a stimulus and another presented immediately after. Historically, repetition suppression has been used to measure the perceptual dimensions coded by a particular region as opposed to internal category structure per se, but recent work suggests that repetition suppression can also contain information about longer term category statistics in a task (Leopold et al. 2006; Panis et al. 2011; Kahn and Aguirre 2012). For example, the level of repetition suppression on a given trial may measure not only the similarity between adjacent stimuli, but also a stimulus' prototypicality or similarity to other stimuli in the task. Psychologically, this adaptation may indicate the retrieval of a norm that subjects compare to current stimuli (Leopold et al. 2006; Panis et al. 2011; Kahn and Aguirre 2012), which is akin to how categorization operates in formal prototype models (e.g., Minda and Smith 2002).

Although repetition suppression has well-established links to similarity and a direct relationship to the underlying neural population firing (e.g., Sawamura et al. 2006) it also has several pitfalls with respect to its interpretation as a measure of internal category structure. First, like other univariate activation measures, repetition suppression can potentially be contaminated by any other psychological variable that is known to impact univariate activation beyond just the representational similarity between 2 adjacent stimuli (e.g., uncertainty, novelty, and salience). Second, in many cases, short-term adaptation differences between stimuli are confounded with longer term adaptation due to stored stimulus/category representations because physically average stimuli more often follow similar stimuli than nonphysically average stimuli (Leopold et al. 2006; Panis et al. 2011; Kahn and Aguirre 2012). Thus, for example, aggregate measures of activation for physically average items can appear to be associated with the highest level of adaptation, even in cases for which activation contains only information about adjacent stimuli.

Because our neural typicality measure is not based on mean activation-level differences between stimuli, it may be more directly interpretable and less susceptible to adjacency effects in studies of longer term internal category structure. Indeed, in a set of follow-up analyses we found that our results remained significant (A hierarchical linear model was used to test the effect of psychological typicality on the neural typicality measure within each ROI while simultaneously controlling for the Euclidean distances between adjacent stimuli. The relationship between psychological and neural typicality remained significant in both ROIs [temporal ROI: $t_{(12)} = 11.25$, $P < 0.001$; early visual ROI: $t_{(12)} = 13.00$, $P < 0.001$].) in both of our ROIs after controlling for the pairwise Euclidean distances between stimuli (see Kahn and Aguirre 2012), suggesting that our neural typicality results are not driven by suppression between adjacent stimuli. Still, in future research, it will be worthwhile to employ designs that use both of these potential measures of internal category structure to better understand their respective advantages and disadvantages. For example, because our intertrial intervals are longer than those employed in repetition suppression tasks (Kahn and Aguirre 2012), it remains to be seen whether our neural typicality measure is unaffected by suppression between adjacent stimuli in all contexts.

Although our neural typicality measure offers a principled and straightforward method for testing an item's relationship to neural category representations, it is still important to continue to test its assumptions in a number of paradigms. By widely testing the measure with both real world and artificial categories, it will help to ensure that our measure is valid in broader contexts and to strengthen its link to psychological category representations. One potential difficulty with strengthening the measure's links to psychological category representations is that what exactly psychological measures of internal structure are measuring is an ongoing field of research in and of itself (e.g., Davis and Love 2010; Kim and Murphy 2011; Voorspoels et al. 2011), and thus measures like typicality ratings can often lead to multiple interpretations within a single task setting. For example, typicality ratings often indicate goodness-of-example, which can diverge from what subjects actually believe is average for a category (Kim and Murphy 2011). Likewise, because typicality ratings are often at least partially correlated with other measures of internal structure such as reaction time (Reaction time alone is insufficient for explaining our results in the present context because it is controlled for in all statistical models.) and probability correct (Posner and Keele 1968; Rosch et al. 1976), they may be interpreted as measures of fluency or certainty. Although our reconstruction results are less ambiguous than typicality ratings (Davis and Love 2010) and suggest that subjects' psychological representations actually are idealized, it is still possible that our fMRI results may indicate that the brain organizes representations around fluently processed category members as opposed to reflecting subjects' actual psychological representations per se. Thus it will be important for future research connecting psychological and neural structure to integrate results over a number of paradigms involving both real world and artificial category structures. By developing a principled measure of neural typicality and showing that neural structure is connected to psychological measures in a manner predicted by computational theories, our results make an important first step in this endeavor.

Mechanisms That Shape Internal Structure

One important question, with respect to our results, concerns the mechanisms by which psychological and neural category representations come to differ from physical stimulus spaces. Cognitive categorization models have a number of options available for emphasizing particular regions or dimensions of a stimulus space. Formally, this amounts to increasing or decreasing the density estimates of particular regions of a stimulus space or the impact of specific stimulus dimensions. One well-studied mechanism that can warp the topography of a physical stimulus space is dimensional selective attention (Nosofsky 1986). Dimensional selective attention allows emphasis on a particular stimulus dimension such that changes

along this dimension are magnified in the psychological or neural space. In the extreme, dimensional selective attention acts as a rule and only values along a single dimension become relevant for a decision (Nosofsky 1991). Recent neuroscience research has found that learning a category can lead to changes in repetition suppression in the occipital and temporal cortices, such that the amount of suppression between 2 stimuli increases along attended dimensions (Folstein et al. 2012). Similarly, in the broader literature, machine learning methods have been used to decode which dimensions of stimuli are attended to based on patterns of activation in early visual and higher level regions (e.g., Haynes and Rees 2005; Kamitani and Tong 2005; MacEvoy and Epstein 2009; Serences et al. 2009). Attentional mechanisms in the PFC that instantiate rule-based strategies (Ashby et al. 1998; Smith et al. 1998; Ashby and Maddox 2005) may contribute to selective attention effects by influencing neural representations in a top-down manner.

In the present context, dimensional selective attention is insufficient for explaining the idealization effect because dimensional selective attention affects an entire dimension uniformally. There is no linear reweighting of the stimulus dimensions that can make idealized items more similar to other items than items that are physically average. Indeed, our model that favored physical averages included dimensional selective attention, and was not able to account for neural or behavioral results. In order for models to predict idealization effects, or any effect in which a subsection of space is emphasized (e.g., rule-plus-exception tasks; Sakamoto and Love 2004, 2006; Davis et al. 2012a, 2012b) additional mechanisms are required. In the present context, within an attentional framework, different regions of the stimulus space (i.e., different exemplars) may be able to be emphasized by exemplar-specific attention mechanisms (Sakamoto et al. 2004; Rodrigues and Murre 2007) that either upweight or downweight the impact of similarity between individual stimuli. Although we know of no studies that have directly explored the neural basis of such exemplar-specific attention mechanisms in categorization, theories about attentional spotlight effects in the processing of visual scenes are interestingly related (e.g., Kastner and Ungerleider 2000; Corchs and Deco 2002). Attention has been found to create a spotlight around salient regions of visual space such that the processing of stimuli close to this location in space is enhanced (not just differences along a specific dimension of visual space; Kastner et al. 1998; Brefczynski and DeYoe 1999). It is conceptually straightforward to predict that the same or similar spotlight mechanisms may affect the topography of stored neural stimulus representations, such that regions of a category space that contain highly idealized category members are enhanced and contribute more to categorization and typicality judgments than exemplars in ambiguous regions of category space.

A second candidate mechanism for idealization effects is error-driven learning mechanisms that adjust category representations to reduce prediction error and confusion between categories (Davis and Love 2010). In these models, category members are simultaneously pulled toward representations/members of their own categories and repelled by members of opposing categories. In brief tasks, the net force on category representations tends to be repelling, and category averages are remembered as more idealized than they actually are (Davis and Love 2010). Unlike attention-based accounts, error-driven learning assumes that category learning can actually change the location of points (or category representations) within the space as opposed to simply emphasizing or de-emphasizing different regions or dimensions. For example, error-driven learning accounts in the present task would predict that the psychological average of tall and short birds would shift apart over the course of learning such that the actual representations are idealized. In terms of neural mechanisms, the error-driven learning account seems less likely to be a viable explanation for the neural effects as actual neuronal changes in regions of early visual cortex happen on a much longer scale than our task (Ghose et al. 2002).

The goal of the present study was not designed to disentangle the psychophysical or neurophysical mechanisms that transform physical stimulus spaces into psychological or neural spaces. However, our finding of relations between neural and psychological category spaces that arise from such transformations paves the way for future research to explore commonalities and differences in the processing mechanisms that lie upstream. In future research aimed at dissociating candidate mechanisms, it will be critical to scan subjects during both learning and test and employ a variety of category structures (e.g., Levering and Kurtz 2006; Davis and Love 2010). Scanning subjects over a longer time period or during multiple training sessions may also help to clarify the underlying neural and psychological mechanisms that shape category representations. It may be the case that, over time, subjects come to better represent the actual physical category structure and thus the amount by which neural and psychological representations are idealized could change. For example, over long time periods, error-driven learning models should converge toward the true category averages whereas attentional effects may never dissipate as long as the same dimensions and stimuli remain salient.

The Role of Perceptual Regions Representing Internal Structure

In addition to testing how cognitive/theoretical mechanisms relate to neural typicality, it will also be useful for future research to begin to delineate the precise role that different perceptual regions play in representing internal structure. Here, we hypothesize that the regions that represent internal structure in any given task will be those that code the differences between stimuli within the categories tested. We predicted early visual cortex would be sensitive to internal structure in the present case because differences between stimuli within a category can be represented with the primitive features early visual cortex is known to code. In other tasks where primitive visual features do not reliably distinguish between categories (or categories based in other modalities), we would expect that early visual cortex would not be recruited to code internal structure.

This view, that it is the features of the categories themselves that determine which perceptual regions will be recruited differs, in some respects, with category-learning theories that emphasize the structure of categories (e.g., verbalizable vs. nonverbalizable rules) and learning demands (e.g., incidental or unsupervised vs. supervised learning) in determining which perceptual regions will be recruited to learn new categories (for review see Ashby and Maddox 2005). One

important finding from this tradition, with respect to the current results, is that early visual cortex activation tends to be more strongly associated with incidental learning tasks (Reber et al. 2003; but see Gureckis et al. 2011) or tasks requiring subjects to learn to discriminate between members and nonmembers of a single category (e.g., A/non-A; Reber et al. 1998a, 1998b; Aizenstein et al. 2000) as opposed to learning to distinguish between members of 2 different categories (Seger et al. 2000; Vogels et al. 2002; Zeithamova et al. 2008; A/B tasks). Our task involves learning to distinguish multiple categories, akin to A/B tasks, and so our finding that early visual cortex is involved with representing category structure may be at odds with theories emphasizing the role of task demands (as opposed to featural qualities) in determining which perceptual regions will be recruited to represent categories. This discrepancy between our current study and previous results and theory may be due to the fact that all previous category-learning studies exploring the role of early visual cortex in category learning have used simple univariate measures of mean activation/deactivation that, unlike our multivariate methods, do not explicitly test whether the relationships between stimuli in the task are coded within any particular region.

Still, it will be important for future research to explicitly test whether and how regions like the early visual cortex represent internal structure in a variety of designs and to explicitly compare A/B and A/non-A tasks. We predict that, in contexts where the differences between stimuli within categories can be represented with primitive visual features, the early visual cortex will be involved with representing internal structure regardless of the task demands. However, which items will be the most psychologically typical will change depending upon the design and task demands, and thus, we expect that which items will be more neurally typical will also change. For example, in unsupervised learning tasks, or many A/non-A tasks, there would be no behavioral benefit to emphasizing extreme or caricatured category members, and thus psychological and neural typicality should favor the true physical category averages (see also, Levering and Kurtz 2006; Davis and Love 2010).

The Plurality of Categorization Research

For much of the history of categorization research (and cognition in general), there has been a trade-off between realism, or ecological studies of real-world categories, and the use of highly controlled novel stimuli. On the one hand, real-world categories are ultimately what categorization researchers wish to generalize to; artificial categories are often much less meaningful and complex, and results from artificial categorization experiments may not always scale up (Murphy 2004). On the other hand, it is not possible to control subjects' prior experience with real world categories, and the physical and psychological features of such categories that subjects use to guide categorization choices are often much less straight-forward than for artificial stimuli. In the present study, we opted for highly controlled artificial stimuli because it was crucial to be able to generate straight-forward predictions for how physical and psychological typicality contributed to neural typicality gradients. It can be difficult to thoroughly disentangle contributions of physical similarity in natural categories because taxonomic categories, such as birds, are often formed precisely because their members share physical and perceptual characteristics (Rosch 1973; Rosch and Mervis 1975). Still, owing to the reciprocal nature of categorization research, it will be important for future research to test the neural typicality measure in more real-world settings. One ambitious idea for future research is to test whether cross-cultural differences in neural typicality gradients reflect well-known psychological differences in the organization of common taxonomic categories. For example, Native American populations tend to organize categories on the basis of ecological principles (e.g., the food chain), whereas western subjects are more likely to organize categories on the basis of perceptual and taxonomic relations (Medin and Atran 2004).

Conclusion

In conclusion, investigating the organization of category representations in the brain offers an important window into the neural basis of cognition. Categories underlie much of everyday cognition and allow us to generalize knowledge or behaviors learned from one object to other related objects. The internal structure or organization of category representations is a key topic in categorization research and has a dramatic impact on categorical inference and behavior. Here, we coined a multivariate neural typicality measure that is able to reveal the internal structure category representations in the brain. Using this measure, we found that the organization of category representations in regions of the temporal and occipital cortex are linked to subjects' psychological category structure, such that, to the extent that stimuli are more typical psychologically, they tend to also be more similar to other members of their category in a neural pattern similarity space. The relationship between pattern similarity and category representation, while still only beginning to be explained, has the potential to uncover general principles that may hold across the brain's representational systems and cognitive domains, suggesting that the brain may be a key source of data for linking cognitive and neurobiological theories of behavior.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Notes

The authors thank Brenda Gregory and Natalie Picchetti for helping in collecting data. They also thank W. Todd Maddox, Michael Mack, and Molly Ireland for helpful comments on a previous version of this manuscript. *Conflict of Interest*: None declared.

Funding

James S. McDonnell Foundation grant to R.A.P.

References

Aizenstein HJ, MacDonald AW, Stenger VA, Nebes RD, Larson JK, Ursu S, Carter CS. 2000. Complementary category learning systems identified using event-related functional MRI. J Cogn Neurosci. 12:977–987.

Aron AR, Shohamy D, Clark J, Myers C, Gluck MA, Poldrack RA. 2004. Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. J Neurophysiol. 92:1144.
Ashby FG, Alfonso-Reese LA. 1995. Categorization as probability density estimation. J Math Psychol. 39:216–233.
Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM. 1998. A neuropsychological theory of multiple systems in category learning. Psychol Rev. 105:442–481.
Ashby FG, Maddox WT. 2005. Human category learning. Annu Rev Psychol. 56:149–178.
Ashby FG, Maddox WT. 1993. Relations between prototype, exemplar, and decision bound models of categorization. J Math Psychol. 37:372–400.
Atran S. 1999. Itzaj Maya folkbiological taxonomy. Folkbiology. 119–203.
Barsalou LW. 1985. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. J Exp Psychol Learn Mem Cogn. 11:629.
Brefczynski JA, DeYoe EA. 1999. A physiological correlate of the 'spotlight'of visual attention. Nat Neurosci. 2:370–374.
Burnett RC, Medin DL, Ross NO, Blok SV. 2005. Ideal is typical. Can J Exp Psychol. 59:3.
Corchs S, Deco G. 2002. Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data. Cereb Cortex. 12:339–348.
Davis T, Love BC. 2010. Memory for category information is idealized through contrast with competing options. Psychol Sci. 21:234.
Davis T, Love BC, Preston AR. 2012a. Learning the exception to the rule: model-based fMRI reveals specialized representations for surprising category members. Cereb Cortex. 22:260–273.
Davis T, Love BC, Preston AR. 2012b. Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fMRI. J Exp Psychol Learn Mem Cogn. 38:821–839.
Diana RA, Yonelinas AP, Ranganath C. 2008. High-resolution multivoxel pattern analysis of category selectivity in the medial temporal lobes. Hippocampus. 18:536–541.
Edelman S. 1998. Representation is representation of similarities. Behav Brain Sci. 21:449–467.
Engel SA, Glover GH, Wandell BA. 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb Cortex. 7:181–192.
Estes WK. 1996. Classification and cognition. New York, USA: Oxford University Press.
Folstein JR, Palmeri TJ, Gauthier I. 2012. Category learning increases discriminability of relevant object dimensions in visual cortex. Cereb Cortex.
Freedman DJ, Riesenhuber M, Poggio T, Miller EK. 2003. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J Neurosci. 23:5235–5246.
Gärdenfors P. 2004. Conceptual spaces: the geometry of thought. Cambridge, MA: MIT Press.
Ghose GM, Yang T, Maunsell JHR. 2002. Physiological correlates of perceptual learning in monkey V1 and V2. J Neurophysiol. 87:1867.
Grindband J, Hirsch J, Ferrera VP. 2006. A neural representation of categorization uncertainty in the human brain. Neuron. 49:757–763.
Gureckis TM, James TW, Nosofsky RM. 2011. Re-evaluating dissociations between implicit and explicit category learning: an event-related fMRI study. J Cogn Neurosci. 23:1697–1709.
Hastie TJ, Tibshirani RJ. 1990. Generalized additive models. London: Chapman & Hall.
Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 293:2425.
Haynes JD, Rees G. 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci. 8:686–691.
Jakel F, Scholkopf B, Wichmann FA. 2009. Does cognitive science need kernels? Trends Cogn Neurosci. 13:381–388.
Jimura K, Poldrack RA. 2011. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. Neuropsychologia.
Kahn DA, Aguirre GK. 2012. Confounding of norm-based and adaptation effects in brain responses. NeuroImage.
Kamitani Y, Tong F. 2005. Decoding the visual and subjective contents of the human brain. Nat Neurosci. 8:679–685.
Kastner S, De Weerd P, Desimone R, Ungerleider LG. 1998. Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. Science. 282:108–111.
Kastner S, Ungerleider LG. 2000. Mechanisms of visual attention in the human cortex. Annu Rev Neurosci. 23:315–341.
Kim S, Murphy GL. 2011. Ideals and category typicality. J Exp Psychol Learn Mem Cogn. 37:1092–1112.
Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. Proc Natl Acad Sci USA. 103:3863–3868.
Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci. 2:1–28.
Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA. 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 60:1126–1141.
Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci. 12:535–540.
Kruschke JK. 1992. ALCOVE: an exemplar-based connectionist model of category learning. Psychol Rev. 99:22.
Leopold DA, Bondar IV, Giese MA. 2006. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature. 442:572–575.
Levering K, Kurtz KJ. 2006. The influence of learning to distinguish categories on graded structure. In: Sun R, Miyake N, editors. Proceedings of the 28th Annual Conference of the Cognitive Science Society; 2006 July 26–29. Austin, TX: Cognitive Science Society. p. 1681–1686.
Liang JC, Wagner AD, Preston AR. 2013. Content representation in the human medial temporal lobe. Cereb Cortex.
Love BC. 2005. Environment and goals jointly direct category acquisition. Curr Dir Psychol Sci. 14:195–199.
Love BC, Gureckis TM. 2007. Models in search of a brain. Cogn Affect Behav Neurosci. 7:90–108.
Love BC, Medin DL, Gureckis TM. 2004. SUSTAIN: a network model of category learning. Psychol Rev. 111:309.
Lynch EB, Coley JD, Medin DL. 2000. Tall is typical: central tendency, ideal dimensions, and graded category structure among tree experts and novices. Mem Cognit. 28:41–50.
MacEvoy SP, Epstein RA. 2009. Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. Curr Biol 19:943–947.
Maddox WT, Ashby FG. 2004. Dissociating explicit and procedural-learning based systems of perceptual category learning. Behav Processes. 66:309–332.
Markman AB, Ross BH. 2003. Category use and category learning. Psychol Bull. 129:592.
Martin A, Wiggs CL, Ungerleider LG, Haxby JV. 1996. Neural correlates of category-specific knowledge. Nature. 379:649–652.
Medin DL, Atran S. 2004. The native mind: biological categorization and reasoning in development and across cultures. Psychol Rev. 111:960.
Miller EK, Freedman DJ, Wallis JD, Miller EK, Freedman DJ, Wallis JD. 2002. The prefrontal cortex: categories, concepts and cognition. Phil Trans R Soc Lond B. 357:1123–1136.
Minda JP, Smith JD. 2002. Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. J Exp Psychol Learn Mem Cogn. 28:275.
Mumford JA, Turner BO, Ashby FG, Poldrack RA. 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage. 59:2636–2643.
Murphy GL. 2004. The big book of concepts. Cambridge, MA: MIT Press.

Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Neurosci. 10:424–430.
Nosofsky RM. 1986. Attention, similarity, and the identification-categorization relationship. J Exp Psychol Gen. 115:39.
Nosofsky RM. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. J Exp Psychol Learn Mem Cogn. 14:700.
Nosofsky RM. 1992. Similarity scaling and cognitive process models. Annu Rev Psychol. 43:25–53.
Nosofsky RM. 1991. Typicality in logically defined categories: exemplar-similarity versus rule instantiation. Mem Cognit. 19:131–150.
O'Toole AJ, Jiang F, Abdi H, Haxby JV. 2005. Partially distributed representations of objects and faces in ventral temporal cortex. J Cogn Neurosci. 17:580–590.
Palmeri TJ, Gauthier I. 2004. Visual object understanding. Nat Rev Neurosci. 5:291–303.
Pan X, Sakagami M. 2012. Category representation and generalization in the prefrontal cortex. Eur J Neurosci. 35:1083–1091.
Panis S, Wagemans J, Op de Beeck HP. 2011. Dynamic norm-based encoding for unfamiliar shapes in human visual cortex. J Cognit Neurosci. 23:1829–1843.
Poldrack RA, Foerde K. 2008. Category learning and the memory systems debate. Neurosci Biobehav Rev. 32.
Posner MI, Keele SW. 1968. On the genesis of abstract ideas. J Exp Psychol. 77:353.
Reber PJ, Gitelman DR, Parrish TB, Mesulam MM. 2003. Dissociating explicit and implicit category knowledge with fMRI. J Cogn Neurosci. 15:574–583.
Reber PJ, Stark CEL, Squire LR. 1998. Cortical areas supporting category learning identified using functional MRI. Proc Natl Acad Sci USA. 95:747–750.
Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. Nat Neurosci. 2:1019–1025.
Rissman J, Gazzaley A, D'Esposito M. 2004. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage. 23:752–763.
Rodriques PM, Murre JM. 2007. Rules-plus-exception tasks: a problem for exemplar models? Psychon Bull Rev. 14:640–646.
Rosch E, Mervis CB. 1975. Family resemblances: studies in the internal structure of categories. Cognit Psychol. 7:573–605.
Rosch E, Simpson C, Miller RS. 1976. Structural bases of typicality effects. J Exp Psychol Hum Percept Perform. 2:491–502.
Rosch EH. 1973. Natural categories. Cognit Psychol. 4:328–350.
Rosseel Y. 2002. Mixture models of categorization. J Math Psychol. 46:178–210.
Sakamoto Y, Love BC. 2004. Schematic influences on category learning and recognition memory. J Exp Psychol Gen. 133:534.
Sakamoto Y, Love BC. 2006. Vancouver, Toronto, Montreal, Austin: enhanced oddball memory through differentiation, not isolation. Psychon Bull Rev. 13:474–479.
Sakamoto Y, Matsuka T, Love BC. 2004. Dimension-wide vs. exemplar-specific attention in category learning and recognition. In Proceedings of the 6th International Conference of Cognitive Modeling. 261–266.
Sawamura H, Orban GA, Vogels R. 2006. Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the FMRI adaptation paradigm. Neuron. 49:307–318.
Seger CA. 2008. How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. Neurosci Biobehav Rev. 32:265–278.
Seger CA, Cincotta CM. 2006. Dynamics of frontal, striatal, and hippocampal systems during rule learning. Cereb Cortex. 16: 1546–1555.
Seger CA, Miller EK. 2010. Category learning in the brain. Annu Rev Neurosci. 33:203–219.
Seger CA, Poldrack RA, Prabhakaran V, Zhao M, Glover GH, Gabrieli JDE. 2000. Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. Neuropsychologia. 38:1316–1324.
Serences JT, Ester EF, Vogel EK, Awh E. 2009. Stimulus-specific delay activity in human primary visual cortex. Psychol Sci. 20:207–214.
Shepard RN. 1987. Toward a universal law of generalization for psychological science. Science. 237:1317.
Smith EE, Patalano AL, Jonides J. 1998. Alternative strategies of categorization. Cognition. 65:167–196.
Spiridon M, Kanwisher N. 2002. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron. 35:1157–1165.
Vanpaemel W, Storms G. 2008. In search of abstraction: the varying abstraction model of categorization. Psychon Bull Rev. 15: 732–749.
Venables WN, Ripley BD. 2002. Modern applied statistics with S. New York Springer.
Vogels R, Sary G, Dupont P, Orban GA. 2002. Human brain regions involved in visual categorization. Neuroimage. 16:401–414.
Voorspoels W, Vanpaemel W, Storms G. 2011. A formal ideal-based account of typicality. Psychon Bull Rev. 18:1006–1014.
Vul E, Harris C, Winkielman P, Pashler H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci. 4:274–290.
Wood SN. 2006. Generalized additive models: an introduction with R. CRC Press.
Xue G, Dong Q, Chen C, Lu Z-L, Mumford JA, Poldrack RA. 2012. Complementary role of frontoparietal activity and cortical pattern similarity in successful episodic memory encoding. Cereb Cortex. doi: 10.1093/cercor/bhs143.
Xue G, Dong Q, Chen C, Lu Z, Mumford JA, Poldrack RA. 2010. Greater neural pattern similarity across repetitions is associated with better memory. Science. 330:97.
Zeithamova D, Maddox WT, Schnyer DM. 2008. Dissociable prototype learning systems: evidence from brain imaging and behavior. J Neurosci. 28:13194–13201.