Expecting the Unexpected: Base-rate Neglect is Driven by Neural Activation of Common Cues

Sean R. O'Bryan*[1], Darrell A. Worthy[2], Evan Livesey[3], Tyler Davis[1]

1 Department of Psychological Sciences, Texas Tech University
2 Department of Psychology, Texas A&M University
3 School of Psychology, University of Sydney

Correspondence to: sean.r.obryan@ttu.edu

ABSTRACT

The inverse base-rate effect (IBRE) describes an apparent irrationality in human decision making whereby people tend to ignore category base rates and choose rarer options when classifying ambiguous stimuli. According to associative learning theories, people choose rare categories for ambiguous stimuli because rare cues draw more attention. Alternatively, inferential theories predict that people choose rare categories because ambiguous stimuli contrast more with well-established rules that predict membership in the common category. In this experiment, we used multi-voxel pattern analysis to decode which features of ambiguous stimuli, common or rare, participants are activating during an fMRI version of the IBRE task. Participants learned to predict four hypothetical diseases based on pairings of face, object, and scene cues. Objects and scenes were diagnostic cues, and predicted either a common or rare disease. Prior to the task, we collected independent localizer scans to characterize activation patterns associated with each of the different visual cues in object-selective cortex. During a test phase, ambiguous trials were presented in which a cue for the rare disease was paired with a cue for the common disease. We show that individuals engage qualitatively distinct neural processes when making rare versus common responses: choosing the rare category involved activation of cues associated with the common category. Consistent with inferential theories of base-rate neglect, our findings suggest that this surprising behavior involves a deliberative mechanism not explained by purely associative models.

KEYWORDS: Attention, Category learning, Irrationality, Reasoning, Representational similarity analysis

1. Introduction

Categorization is a critical human capacity, enabling us to organize what we have learned in a way that generalizes to new experiences. Knowledge about the relative frequency of different categories in the environment can provide a valuable source of evidence when predicting category membership. Consider a doctor tasked with diagnosing a patient who presents with two symptoms – one associated with a common disease, and the other associated with a rare disease. When faced with such an ambiguous case where a patient exhibits symptoms of multiple diseases, it would be rational to take into account the disease base rates and choose the most common diagnosis. Yet, research shows that base-rate information is applied inconsistently in both medical decision making (Casscells, Schoenberger, and Graboy, 1978; Bravata, 2000) and problem solving contexts (Tversky and Kahneman, 1974).

Cases in which base-rate information is improperly used can be found in the laboratory, even in simplistic learning and decision making tasks. Here we focus on one such task that has been reliably shown to produce an inverse-base rate effect (IBRE), where participants' response preferences directly contradict the category base rates when making predictions about an ambiguous combination of stimuli (Medin and Edelson, 1988; Kruschke, 1996). In this task, participants are trained to predict which of four possible diseases a participant has based on combinations of verbal or visual cues. On each trial, participants are presented with two cues: one an imperfect predictor *I*, and the other either a common perfect predictor *PC* or rare perfect predictor *PR*. Each perfect predictor has a deterministic relationship with a disease, whereas the imperfect predictor is associated with multiple diseases. The only distinction between

PC and *PR* is the disease base rates, with common diseases appearing three times more often than the rare ones throughout training. After learning, the key IBRE finding is that participants tend to ignore the base rates and choose the rare category when confronted with an ambiguous stimulus that contains both of the previously perfect common and rare predictors (*PC.PR*).

A recurring debate surrounding the cognitive basis of the IBRE is whether it is the byproduct of high-level reasoning processes (Winman, Wennerholm, and Juslin, 2003; Winman et al., 2005) or a fundamental associative mechanism (Kruschke, 2001a; Kruschke, 2003; Johansen, Fouquet, and Shanks, 2007; Lamberts and Kent, 2007). In the present study, we sought to mechanistically test these two theories of how and why the brain misuses category base rates by using functional magnetic resonance imaging (fMRI) to decode attentional processes underlying the brain's processing of ambiguous information.

Most efforts to model the IBRE have focused on the role of attention in learning, and how fundamental associative processes can produce this seemingly irrational behavior. Thus far, the most successful associative models for accommodating the IBRE were formulated by Kruschke (1996; 2001b). According to associative accounts, the attentional mechanism proposed to explain the IBRE is grounded in asymmetrical cue representations that form early during learning. Because the pairing of the imperfectly predictive cue and the common perfectly predictive cue (*I.PC)* occurs more frequently, participants first learn that *I* + *PC* → *Common*. Consequently, when encountering the less-frequent pairings of the imperfectly predictive cue and the rare perfectly predictive cue (*I.PR*), individuals shift their attention to *PR* and away from *I*,

because *I* has already been associated with the common disease. The stronger tendency to choose the rare category when confronted with the ambiguous pairing of the perfectly predictive common and rare cues (*PC.PR)* is then explained by the stronger association that forms between *PR* and the rare outcome due to learned selective attention (Medin and Bettger, 1991; Kruschke, 1996).

Despite the consistency between the results of cue learning tasks and predictions from associative learning models, disagreement remains as to whether the IBRE is the result of a fundamental associative learning mechanism, or whether higher-order inferential processes contribute to the surprising pattern of results (Juslin, Wennerholm, and Winman, 2001; Winman et al., 2005). Like the associative accounts, an inferential account may predict that people will learn the more common disease earlier in the task and the most strongly. However, inferential accounts differ in terms of why people chose the rare stimulus on the ambiguous *PC.PR* trials. From an inferential account, because *I* + *PC* → *Common* is so well learned, *PC.PR* is more unexpected for the common category, which leads participants to infer that such a combination would be inconsistent with the common category, and thus must come from the rare category (Juslin et al., 2001). Specifically, according to inferential theories, participants choose the rare category because the presence of the rare cue creates a stronger sense of mismatch between the presented stimuli and their memories of the common category.

Although associative and inferential theories posit very different mechanisms as the basis of the IBRE, they have remained unresolvable at the level of behavior thus far. One possibility to resolve this debate is to begin to look to the brain to provide additional data on the underlying processes occurring during IBRE tasks that remain

hidden to behavior. As a first pass, the goal of distinguishing between the neural regions associated with high-level inferential processes, such as hypothesis testing, rule switching, and problem solving and lower-level associative learning have been one of the main foci of neurobiological category learning research (Ashby and Maddox, 2005; Smith and Grossman, 2008; Seger and Miller, 2010). For example, rule-based processes, akin to those underlying inferential theories, are thought to depend upon cortico-striatal loops involving the PFC and head of the caudate nucleus, whereas simple visual associations may be learned using visual and procedural loops connecting the tail of the caudate with visual regions and motor outputs.

Despite well-established differences in the neural systems engaged for rule-based and associative tasks (Nomura et al., 2007; Schnyer et al., 2009), relying on such differences to infer whether rule-based or associative mechanisms are engaged in a particular task is much more tenuous. For one, to do so commits an affirming the consequent or “reverse inference” fallacy (Poldrack, 2006); many brain regions activate for a variety of cognitive functions, and thus activation alone does not permit inferences about the underlying process. Indeed, in a recent study examining a patterning task that is largely agreed upon to differentially tap associative and inferential mechanisms, it was not possible to cleanly infer that separable systems were involved based on the activation patterns alone (Milton et al., 2016).

Although classic localization-based fMRI would not be able to firmly establish the mechanisms underlying the IBRE and resolve the debate between inferential and associative theories, recent advances in representational multi-voxel pattern analysis (MVPA) may (Norman et al., 2006; Kriegeskorte, Mur, and Bandettini, 2008; Davis and

Poldrack, 2013). Instead of focusing on localization of cognitive function, MVPA allows researchers to map information contained in fMRI patterns onto experimental and psychological states. In the simplest form, MVPA can allow researchers to perform “mind-reading” and decode the perceptual objects participants are looking at or retrieving from memory (Rissman and Wagner, 2012; Haxby, Connolly, and Guntupalli, 2014).

Recently, MVPA decoding approaches have become useful for testing how top-down and bottom-up attentional mechanisms impact activation of stimulus representations in perceptual processing regions (Gazzaley and Nobre, 2011). For example, MVPA has been used to test mechanistic theories of how attention modulates activation of predictive information being held in working memory (Lewis-Peacock and Postle, 2008; Lewis-Peacock et al., 2012) and how bottom-up attentional orienting to visual stimulus features (Leong et al., 2017) manifests in enhanced neural activation of these features in ventral temporal cortex. Applications of these attentional findings to category learning have shown that, as a result of learning, stimulus dimensions that are predictive of category membership become more prominent in activation patterns in ventral temporal cortex (Mack et al., 2013) and stored stimulus representations in the hippocampus come to reflect these attention-based changes in neural stimulus representations (Mack et al., 2016).

Results from previous studies on dimensional selective attention in category learning suggest that MVPA could be used to test which theories, associative or inferential, provide the best account of IBRE. Specifically, we will test whether activation patterns elicited for ambiguous stimuli (combinations of perfectly predictive

rare and common cues; *PC.PR*) are consistent with participants focusing more on the rare or common information.

In order to test how information associated with different cues is attended to and activated during the current task, we used real world categories – faces, scenes, and objects as the predictive cues in our task. Faces, scenes, and objects have a well-defined representational topography across the cortex, and are readily decodable even with very basic similarity-based analysis (Haxby et al., 2001). Thus we expected these real-world categories would allow maximal power for discriminating subtle changes in attentional weighting in the present task. Further, by localizing regions sensitive to these cues prior to engaging participants in the experimental task, we were able to create an independent neural similarity space that is unbiased with respect to any learning-dependent salience effects, and thus allows stronger representational conclusions (Davis et al., 2014a). Specifically, we can examine how expression of face, scene, and object activation patterns during the key ambiguous test trials relates to those predicted by associative and inferential accounts.

Because associative and inferential theories differ in terms of whether participants are mainly considering the rare cues (associative theories) or common cues (inferential theories) on ambiguous test trials where both predictors are present (*PC.PR*), it is possible to use MVPA to distinguish between the two theories by examining the activation of rare and common cues during these ambiguous test trials. Cognitively, associative models assume that participants orient their attention to the perfectly predictive rare cue (*PR*) on ambiguous trials by virtue of its perceived distinctiveness for predicting the rare disease (Kruschke, 2001b). Thus, if the IBRE is

attributable to the associative mechanism of learned selective attention, ambiguous test trials should be accompanied by stronger activation of patterns associated with the rare cue (*PR*).

On the other hand, inferential models predict that participants focus on the discrepancy between the ambiguous combination of the common and rare cues (*PC.PR*) and their expectations of the common category (Juslin et al., 2001). From this perspective, the participants should be focusing more on the information associated with the common category when eliminating the ambiguous combination as a possible member of the common category. As such, if a form of inference were responsible for the IBRE, one would expect greater activation of patterns associated with the common cue (*PC*) relative to those associated with the rare cue (*PR*). Unlike associative effects that are thought to apply consistently across responses during the test phase, the use of inference is thought to occur primarily on ambiguous trials receiving rare responses (Juslin et al., 2001). The representational predictions for both theories are summarized in Table 1.

Table 1. Summary of the representational predictions made by associative and inferential accounts.

	Learning Trials	**Learning-test Relationship**	**Ambiguous Test Trials**
Associative Account	Attention to predictive cues *PC* and *PR* is learned quickly; *PR* is attended to more strongly on *I.PR* trials than *PC* is attended to on *I.PC* trials	Individual differences in attention to cue *PR* predict base-rate neglect	Rare cue *PR* is more strongly attended than common cue *PC*; response-independent
Inferential Account	Inference rules are learned quickly; stimuli match rules and elicit direct induction for common rule on *I.PC* trials and for rare rule on *I.PR* trials	Individual differences in activation of the common rule predict base-rate neglect	Common rule is more often active than rare rule; this difference is larger on trials receiving a rare response vs trials receiving a common response

In addition to our MVPA predictions for how pattern similarity will differ according to associative and inferential accounts of the IBRE, we expect univariate differences in brain regions involved with computing uncertainty and decision evidence signals between novel and well-learned stimuli. During the test phase, participants are exposed to cue pairings that have been previously learned or are presented in novel combinations. Novel stimuli should be associated with higher uncertainty, whereas well-learned stimuli should be associated with more certainty. Recent neuroimaging studies suggest that the lateral and dorsal PFC track categorization decisions with higher uncertainty (DeGutis and D'Esposito, 2007; Seger et al., 2015; Davis, Goldwater, and Giron, 2017) whereas the ventromedial prefrontal cortex (vmPFC) tracks decisions with high decision evidence (Lebreton et al., 2015; Davis et al., 2017).

To foreshadow our results, consistent with inferential accounts of the IBRE, our analysis revealed stronger activation of neural patterns associated with common cues on the ambiguous *PC.PR* trials that were categorized as a rare disease. No differences

in cue activation were observed when responses were consistent with category base-rates on these ambiguous trials. Further, we show that the representational strength associated with common predictors for each participant during the learning phase is positively associated with individual differences in the tendency to ignore category base-rates at test. These results thus not only help to clarify the mechanisms underlying base-rate neglect, but also highlight the power of fMRI-based decoding for elucidating cognitive processes that have remained unresolvable at the level of behavior for decades.

2. Materials and Methods

2.1. Participants

Twenty-four healthy right-handed volunteers (age range 18 – 58, mean age ± SEM = 25.08 ± 1.62, 13 women) were recruited through online newsletters and flyers posted at Texas Tech University. Two participants were excluded from the final analysis: one due to excessive head motion and one for falling asleep in the scanner. All subjects provided informed consent prior to participation, and were compensated $35 for a 1.5-hour session. The study protocol was approved by the Texas Tech University Human Research Protection Program.

2.2. Experimental Paradigm

Participants completed a category learning task based on a structure employed by several previous studies (Medin and Edelson, 1988; Kruschke, 1996) with a premise of predicting disease outcomes in hypothetical patients. The category structure (Figure

1) included four diseases overall, with two common diseases that were encountered at a rate of 3:1 compared with the rare diseases. Faces, objects, and scenes were used as cues to maximize the ability to detect dimensional selective attention via MVPA. Accordingly, participants were instructed that they would be learning to diagnose different patients based on people they have interacted with, places they have been, and objects that they recently used. Faces were always imperfect predictors, in that each face was associated with two different disease outcomes. Objects and scenes had a deterministic relationship with the disease outcomes, in that each distinct image was associated with only one disease. The face, object, and scene images used were black-and-white squares presented on a white background with black text. Learning occurred over three scanning runs, with each run presenting four blocks of the stimulus set, resulting in 32 trials per run and 96 learning trials over the whole training phase. Individual trials involved presenting an image pair side-by-side on the projection screen and prompting participants to respond to the question, “Disease 1, 2, 3, or 4?” with a button press within 3 s. Feedback was then provided for 1.75 s, indicating a correct or incorrect response accompanied by the correct disease label. The image pair remained on the screen during feedback. Variable fixation periods drawn from truncated exponentials (mean = 3 s) separated stimulus presentation from feedback, and feedback from the next trial (mean = 2 s). Within-pair stimulus position (left or right) was randomized on each trial, and presentation order of cue pairs was randomized within each block for every participant. Figure 1 depicts the progression of a trial during learning. Participants were randomly assigned to one of two conditions to balance

which images were presented together during training and test, and disease labels were randomized across participants.

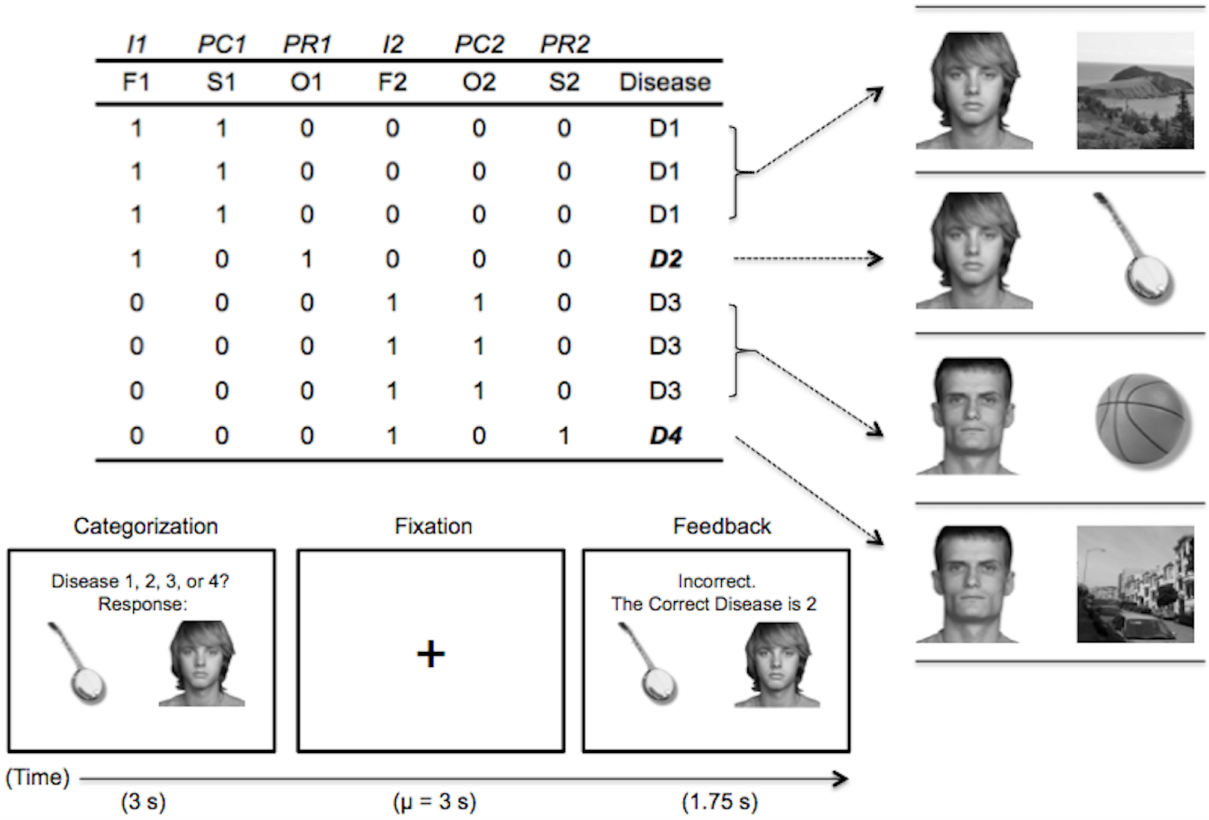

Figure 1. Abstract task design and an example trial. In the headings, I = imperfect predictor, PC = common perfect predictor, PR = rare perfect predictor. The second row refers to the specific stimuli used for each cue: F = face, S = scene, O = object. Each following row corresponds to a learning trial, with a "1" indicating the presence of the cue and "0" indicating its absence.

Following the learning phase, participants engaged in three runs of a test phase where they were asked to classify novel combinations of the previously trained cues as predicting disease 1, 2, 3, or 4. As in learning, participants were given a maximum of 3 s to make a response on each trial. Variable fixation (mean = 3 s) separated each trial. Feedback was not provided following responses during test. The test set included the cue pairs from the original training set occurring at the previously-encountered 3:1

ratio, each cue individually, two presentations of each possible *PC.PR* pair, and the perfect predictors paired with a previously unpaired imperfect predictor (see Table 2, Results). In all, there were 18 unique test items and 26 total items in the test set. Presentation order of the test items was randomized for each of the three runs, with participants rating two test sets per run, for a total of 156 items throughout the test phase.

Prior to beginning the disease prediction task, participants completed two functional localizer scans. The fMRI data collected in these preliminary scans was used to both identify the location of peak BOLD responses to each visual stimulus categories (faces, objects, and scenes) for each participant and estimate distributed patterns of neural activation that were characteristic of the three stimulus categories. Each trial of the localizer task involved presenting a real-world face, object, or scene individually on the screen and asking subjects to make a button press corresponding to the appropriate item category within 2.5 s. Variable fixation lengths drawn from a truncated exponential (mean = 3 s) separated each trial. Over the duration of the localizer phase, subjects categorized 38 examples of each stimulus type. The black-and-white images used during the localizer runs were presented in a random order and did not include any of the stimuli used for the experimental task.

2.3. Image Acquisition

Imaging data were acquired on a 3.0 T Siemens Skyra MRI scanner at the Texas Tech Neuroimaging Institute. Structural images were acquired in the sagittal plane using MPRAGE whole-brain anatomical scans (TR = 1.9 s; TE = 2.44 ms; θ = 9°; FOV

= 250 x 250 mm; matrix = 256 x 256 mm; slice thickness = 1.0 mm, slices = 192). Functional images were acquired using a single-shot T2*-weighted gradient echo EPI sequence (TR = 2.5 s; TE = 25 ms; θ = 75°; FOV= 192 x 192 mm; matrix = 64 x 64; slice thickness = 3 mm).

2.4. fMRI Analysis and Preprocessing

Functional data were preprocessed and analyzed using FSL (www.fmrib.ox.ac.uk/fsl). Anatomical images were preprocessed using Freesurfer (autorecon1). Time series were skull stripped, motion corrected, prewhitened, and high-pass filtered (cutoff: 60 s). For univariate analysis, functional data were spatially smoothed using a 6 mm FWHM Gaussian kernel. No smoothing was performed on functional data used for the representational similarity analysis (RSA; Kriegeskorte et al., 2008). First-level statistical maps were registered to the Montreal Neurological Institute (MNI)-152 template using boundary-based registration to align the functional image to the structural image, and 12 df to align the structural image to the MNI-152 template.

Three-level statistical analysis of the functional data was carried out using FSL's FEAT. At level one, within-run associations between task regressors and functional time series were examined, convolving each regressor with a double-gamma hemodynamic response function. Nuisance regressors were included motion parameters and their temporal derivatives. Additional regressors were included to censor (scrub) volumes that exceeded a framewise displacement of 0.9 mm (Siegel et al., 2014). Additionally, prewhitening was performed at level 1 to control for temporal

autocorrelation in the hemodynamic response. At level 2, within-run parameter estimates for task variables were averaged for each subject using a fixed effects model. Group-level statistical maps were estimated using non-parametric tests in FSL's Randomise and *a priori* thresholding settings. These settings included mass-based cluster thresholding to correct for multiple comparisons at $p < .05$, a cluster-forming *t*-critical value of 2.52 (critical value of *t* for df = 21 and alpha = 0.01, one-tailed), and 6 mm variance smoothing.

Nine task regressors were included in a level-1 model for each of the unique trial types during test. This included onsets for *I*, *PC*, *PR*, *PC.PR* trials receiving a common response, *PC.PR* trials receiving a rare response, previously trained common pairs *I.PC*, previously trained rare pairs *I.PR*, and novel pairs *I.PCo and I.PRo*. Nuisance regressors were included for trial onsets where participants failed to make a response within the cutoff time. Two contrasts of interest based on this model were *PC.PR* trials with a rare response > *PC.PR* with a common response, and novel combination *I.PCo*/ *I.PRo* trials > old *I.PC*/ *I.PR* trials that were pretrained, as different theories of the IBRE have disagreed over the existence of response-dependent strategy use and the role of novelty/uncertainty in the effect. No significant activations were observed for rare versus common responses on *PC.PR* trials, thus this contrast is not discussed further in the paper.

2.5. Multivariate fMRI Analysis

RSA was conducted using the PyMVPA toolbox (Hanke et al., 2009) and custom Python routines. To obtain trial-by-trial estimates of the hemodynamic response, we

computed a β-map (Rissman, Gazzaley, and D'Esposito, 2004) for each stimulus onset using an LS-A procedure (Mumford et al., 2012), simultaneously modeling the trials of interest as separate regressors in a GLM. The same motion regressors were included in the LS-A model as the primary univariate models described above. Because the composition of noise in the scanning environment can determine which method of estimation is optimal (Mumford, Davis, and Poldrack, 2014; Abdulrahman and Henson, 2015), independent tests of classification accuracy using a localizer task identical to that used in our experiment were conducted to reveal that that the LS-A approach was preferable to LS-S (Turner et al., 2012) for the present analysis. The estimated neural activation patterns for each onset were then registered to standard space.

The use of localizer-based multi-voxel pattern analysis enabled us to measure representational similarity to individual cues within the pairs encountered by participants during learning and test (Kriegeskorte et al., 2008). If distributed activation patterns on multi-cue trials more closely resemble those characteristic of one visual image category versus the other, it would suggest that participants are selectively attending to that information. Critically, we sought to use this technique to measure attention to rare versus common cues on the ambiguous *PC.PR* trials that are known to elicit non-rational rare response patterns. To obtain dimensional selective attention predictions, we computed similarities between the patterns for each trial in the experimental paradigm and the patterns computed for face-, object-, and scene-only trials from the independent localizer. The β-series used to compute the multi-voxel patterns was spatially localized in visual stimulus category-specific ROIs by creating 6-mm spheres around subjects' peak activation within anatomically defined regions in the

Harvard-Oxford Atlas associated with category selectivity (objects: left inferior posterior temporal gyrus; scenes: bilateral parahippocampal gyrus; Ishai et al., 1999; Lewis-Peacock and Postle, 2008; Lewis-Peacock et al., 2012; Grill-Spector and Weiner, 2014). The anatomical distribution of the aforementioned spheres across subjects is depicted in Supplementary Figure A1. For the learning phase, the representational analysis was restricted to a functional localizer mask in ventral visual stream that encompassed face-, object-, and scene-selective cortical regions (Supplementary Figure A2). For both learning and test, a Pearson correlation was used to compute 1 – *r* correlation distances between trials on the disease prediction task with face, object, and scene patterns within each subject. Each subject-level correlation map was transformed using Fisher's *Z* and aggregated over trial type for statistical comparison. Descriptive and test statistics for both the MVPA and behavioral analysis were calculated using R (https://www.r-project.org/).

2.5.1. Searchlight Analysis

Because our primary MVPA analysis focused on testing differences in neural similarity to the object and scene cues that comprised the ambiguous test stimuli in ventral occipitotemporal cortex, a question of interest is whether any regions beyond our category-selective ROIs contain representations consistent with either associative or inferential theories of base-rate neglect. Accordingly, we conducted a series of RSA searchlight analyses to determine where pattern similarity tracked the predictions of both accounts. Regions representing information consistent with the associative account would be expected show greater similarity to the rare cue on ambiguous trials

that receive a rare response, whereas regions representing information consistent with the inferential account would be expected show greater similarity to the common cue on these trials.

A whole-brain searchlight algorithm (Kriegeskorte, Goebel, and Bandettini, 2006) using a 3-voxel radius was thus employed to localize regions where patterns were more similar to rare versus common cues, and vice versa, on the key *PC.PR* test trials. Additionally, we conducted two separate searchlights restricted to the object and scene spaces, respectively, to control for any potential contributions of category salience to the resulting correlation maps in the combined searchlight. For each searchlight, the resulting subject-level correlation maps were computed using the procedure described in section 2.5., while group-level statistical maps were created using permutation tests in FSL's Randomise following the thresholding settings described in section 2.4. No significant group-level clusters were observed for either the rare > common or common > rare contrasts for any of the three searchlights after correcting for multiple comparisons, and as such, are not discussed further in the paper.

3. Results

3.1. Behavioral Results

Learning curves over the 12 learning blocks for common and rare disease item pairs are shown in **Figure 2**. All subjects reached greater than 90% accuracy over the last 4 blocks (M = 98.1%, SD = 2.4%, range = 93.5 – 100%). Mean choice performance in the first block was above chance (25%) for both common (M = 63.6%) and rare (M = 43.2%) cue pairs. Consistent with associative accounts that rely on differences in early

learning, a mixed effects ANOVA revealed a significant block by cue type interaction, F (1, 21) = 9.87, p = .005: the common diseases were learned more quickly than the rare diseases, with prediction accuracy for the common and rare cue pairs becoming comparable by the fifth learning block.

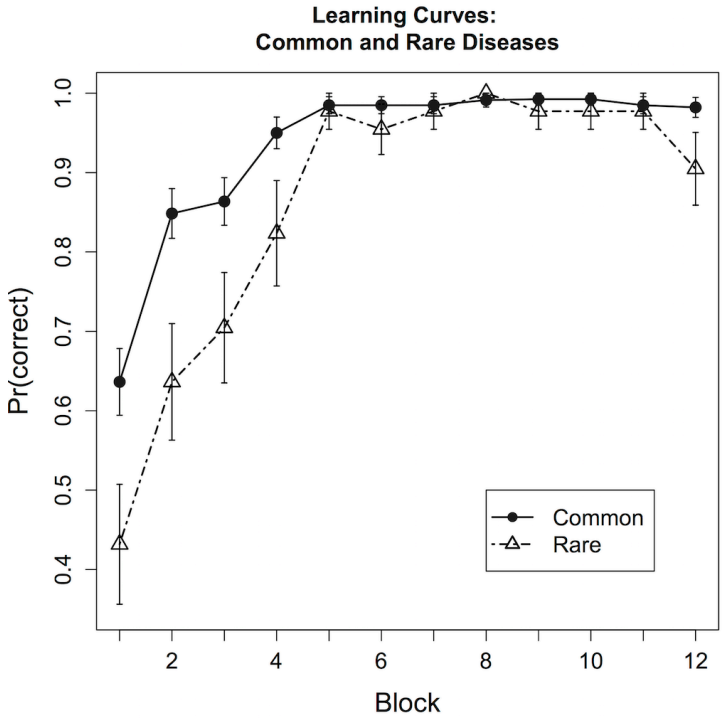

Figure 2. Proportions correct for common and rare disease predictions over the 12 blocks of the training phase (mean ± SEM). Note that for both disease base-rates, performance was well above chance (25%) in the first block.

Results for the test phase are summarized in Table 2. Reported values are aggregated for each trial type across the associated diseases. To determine whether subjects neglected base rates on the critical trials, choice proportions for *PC.PR* and *PC.PRo* (*PRo* being the *other* rare cue that was not paired with *PC* during training) were compared to those for the imperfect predictor *I* (Lamberts and Kent, 2007). Consistent with previous findings, ambiguous perfect predictor pairings were categorized as members of the rare category significantly more often (M = 49.9%) than the imperfect predictors (M = 29.5%), t (21) = 4.10, p < .001, suggesting that participants largely ignored the category base rates on the trials of interest. On

untrained *I.PRo* trials, subjects chose the direct rare disease significantly more often than all other options (*M* = 68.0%), *t* (21) = 2.36, *p* = .028, lending further support to the idea that rare predictors were viewed as distinctive. The perfect predictors *PC* and *PR* produced the expected strong response patterns in favor of the disease they predicted when presented alone, as did the previously-trained *I.PC* and *I.PR* trials.

Table 2. Item combinations and choice proportions for the test phase.

Trial Type	Common	Rare	Other
I	0.608	0.295	0.097
PC	0.878	0.038	0.084
PR	0.008	0.910	0.082
PC+PR	0.435	0.474	0.091
PC+PRo	0.433	0.523	0.044
I+PC	0.979	0.013	0.008
I+PR	0.052	0.929	0.019
I+PCo	0.619	--	0.381
I+PRo	--	0.680	0.320

In addition to response probabilities, we tested whether reaction times differed on *PC.PR* test trials depending on whether a rare or common response was made. On these trials of interest, it was found that RTs were considerably slower when subjects made a rare response (*M* = 1.47 s) than when they made a common response (*M* = 1.27 s), *t* (21) = 5.73, *p* < .001. The observation of slowed RTs on ambiguous trials receiving rare responses indicates that rare selections may involve more conscious

deliberation relative to common selections, which may be the result of simple feature-based responding.

3.2. MVPA Results: Learning

The task structure (Figure 1) demanded that participants learn to selectively attend to objects or scene cues in order to accurately predict each disease outcome. Given the behavioral result that participants were able to categorize both diseases at an above-chance level during the first learning block, it was expected that representational changes associated with attention to these predictive cues would emerge early in learning. Additionally, we predicted that once a high level of performance was attained on the task, attention to the predictive cues would remain stable due to the relative absence of prediction errors to motivate further learning (Hollerman and Shultz, 1998). To estimate neural activity indicative of attention to each cue category, we computed correlation distances between multi-voxel activation patterns on multiple-cue learning trials and patterns elicited by faces, objects, and scenes individually from the independent localizer task. It was hypothesized that representational similarity to each visual category would be strongest for perfectly predictive cues, moderate for the present but non-predictive cues (faces) and lowest for the non-present cues. Figure 3 depicts the time-course of pattern similarity for predictive, non-predictive, and non-present cues during learning for both rare and common disease pairs. Lower values indicate less correlation distance or greater similarity. As predicted, a one-way within-subjects ANOVA collapsed across trial type revealed that neural similarity to the visual category was the strongest for perfectly

predictive cues (M = -0.071), followed by the non-predictive but present cues (M = 0.043) and the non-present cues (M = 0.139), F (2, 42) = 41.4, p < .001. This finding, whereby activation patterns elicited for stimuli during learning show stronger expression (or activation) of information associated with the most predictive cues is consistent with previous studies on dimensional selective attention in category learning (Mack et al., 2013; 2016; Leong et al., 2017).

In order to control for the possibility that more frequent incorrect responses (and consequently, negative feedback) on the rare disease trials contributed to the observed similarity patterns, we conducted the same one-way ANOVA as above but restricted the analysis to trials receiving a correct response. The results were qualitatively identical to the analyses combining incorrect and correct trials, F (2, 42) = 35.9, p < .001, suggesting that stronger activation of patterns associated with the perfectly predictive cues *PR* and *PC* is not attributable to differences in feedback processing.

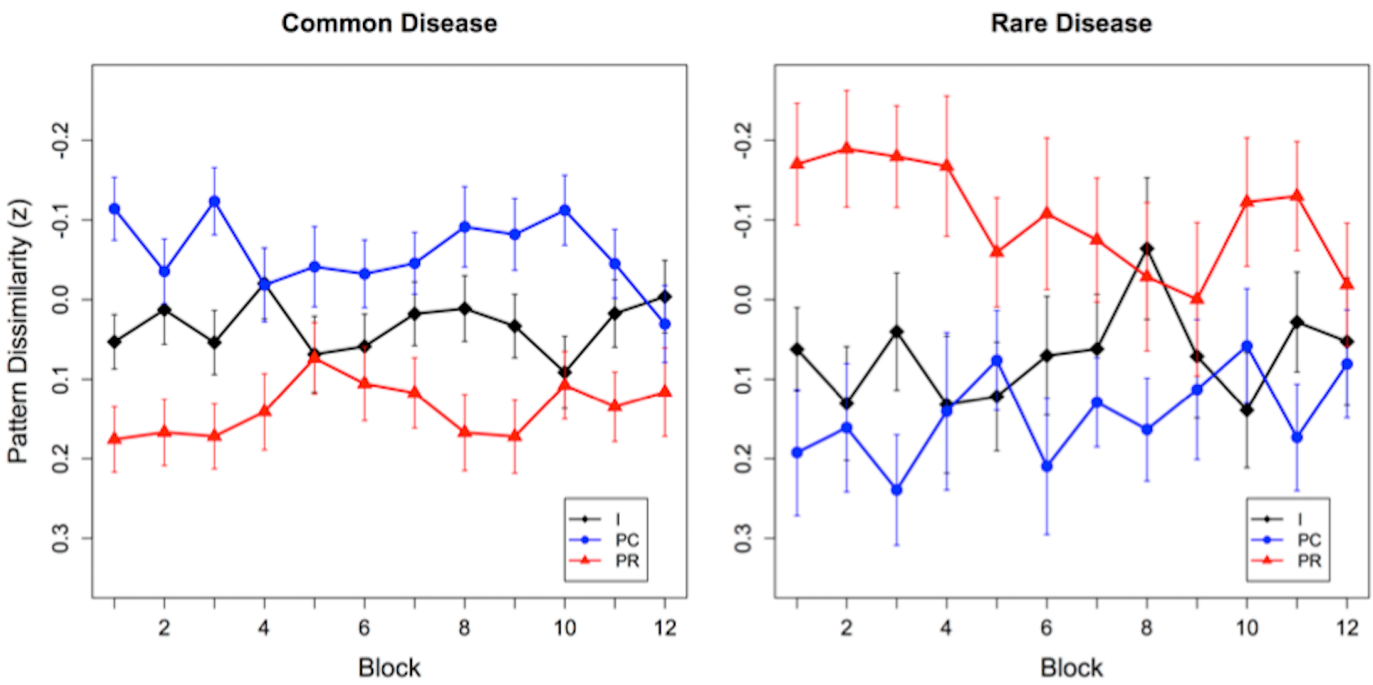

Figure 3. Time-course of multivariate fMRI pattern similarities over the twelve learning blocks (mean ± SEM). The left panel is for trials predictive of a common disease, and the right for trials predictive of a rare disease. For both graphs, the y-axis represents *z*-scored dissimilarity values; the axis is flipped to aid interpretation (more negative = greater neural similarity). 'I' (black) corresponds to imperfect predictors, 'PC' (blue) corresponds to common perfect predictors, and 'PR' (red) corresponds to rare perfect predictors.

3.3. Individual Differences in Behavior Associated with Cue Activation

The presence of a relationship between cue activation during learning and participants' tendencies to engage in base-rate neglect at test may lend support to either the inferential or associative theories of IBRE according to whether activation of common or rare cues during learning predicts test behavior. Based on the predictions of associative models, the degree of attention allocated to rare predictors during learning should relate to test behavior, with greater representational strength for *PR* cues being associated with a stronger tendency to choose the rare category on ambiguous *PC.PR* trials. Conversely, inferential accounts would predict that relatively stronger representations of the common predictor *PC* during learning would lead to an

increased tendency to choose the rare category on *PC.PR* trials at the subject level, as more robust associations between *PC* and the common category should result in a greater sense of mismatch between the ambiguous stimuli and this well-established rule.

Accordingly, we computed Pearson correlations to test whether the representational strength of common or rare cues during learning was associated with the extent to which participants engaged in base-rate neglect at test. Base-rate neglect values for each participant were defined as the proportion of rare responses made on ambiguous *PC.PR* and *PC.PRo* test trials. We quantified the representational strengths for common and rare cues during learning by averaging the neural similarities to common cues *PC* on trials that predicted a common disease, and averaging the neural similarities to rare cues *PR* on trials that predicted a rare disease over all 3 scanning runs (12 blocks) of the learning phase. Consistent with the predictions of inferential theories, we found that greater activation of neural patterns associated with the common predictors during learning was associated with a higher proportion of rare choices on key test trials, $r = -.549$, $t\ (20) = -2.94$, $p = .008$. Alternatively, we found no relationship between activation of neural patterns association with the rare cues *PR* during learning and choice proportions on IBRE trials, $r = -.156$, $t\ (20) = -.705$, $p = .489$. The association between neural similarity to *PC* and test behavior is depicted in Figure 4.

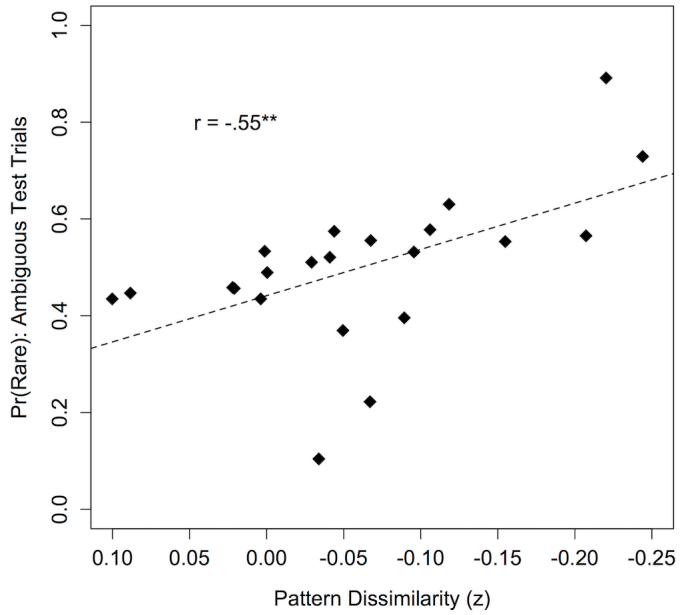

Figure 4. Association between activation of common predictors during learning and individual differences in base-rate neglect. The y-axis represents the proportion of rare responses made on ambiguous *PC.PR* trials during the test phase. The x-axis represents *z*-scored dissimilarity values for the common predictor *PC* over the course of learning for each subject; the axis is flipped to aid interpretation (more negative = greater neural similarity). '**' = $p < 0.01$.

3.4. MVPA Results: Test

The primary research question driving our investigation concerned the distribution of attention to common and rare cues on the ambiguous test trials that produce the IBRE. As stated above, to measure dimensional selective attention, we used MVPA and an independent localizer task where subjects categorized face, object, and scene images one at a time. We then computed similarities between activation patterns elicited during the test phase to those characteristic of each image type. The similarity-based attention measure accurately predicted attention to control items during test; similarity values for objects were significantly greater on object-only (combined across O1, O2, and O1+O2) trials than scene-only trials (combined across S1, S2, and S1+S2), t (21) = -2.82, p = .010, and vice versa, t (21) = -5.61, p < .001. To test whether patterns of attention were independent of behavior, similarity values for rare and common cues were partitioned on a trial-by-trial basis according to whether

the subject responded with a rare or common disease. Findings for the RSA test analysis are depicted in Figure 5. A mixed analysis of variance revealed a significant interaction between response and similarity to chosen/unchosen cues, F (1, 20) = 5.17, p = .034. For IBRE trials receiving a rare response, there was a strong advantage of similarity to the common cue (M = -0.164) over the rare cue (M = 0.077); t (21) = -3.76, p = .001. Alternatively, no differences in pattern similarity were detected for common and rare cues on IBRE trials receiving a common response, t (21) = 0.226, p = .823. Mean similarity values (collapsed across common and rare similarity) did not differ between trial-level responses, t (21) = -0.388, p = .702.

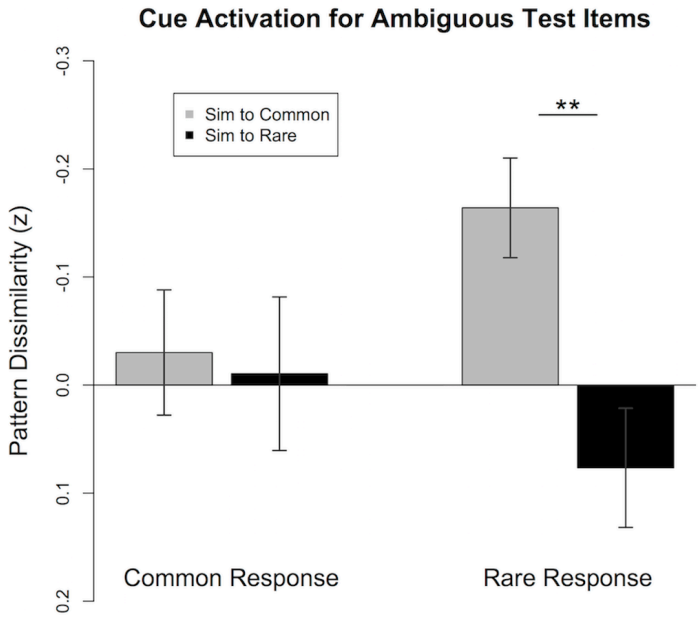

Figure 5. Results from the representational similarity analysis for ambiguous test trials *PC.PR*. The pair of bars on the left reflects activation patterns on trials receiving common responses, while the pair of bars on the right reflects activation patterns on trials where the participant made a rare response. Grey bars depict neural similarity to the common cue, and black bars depict neural similarity to the rare cue. The y-axis represents *z*-scored dissimilarity values (mean ± SEM); the axis is flipped to aid interpretation (more negative = greater neural similarity). '**' = p < 0.01.

To examine potential between-subject effects, a categorical predictor was assigned to each subject to reflect whether their mean choice proportions were consistent with the IBRE on key test trials (pr(Rare) > .50: n = 12, pr(Rare) < .50: n =

10). The interaction between neural similarity and response did not differ according to subjects' mean choice proportions, F (1, 20) = 1.31, p = .266, suggesting that the significant trial-level interaction was consistent across participants. In sum, our RSA results for the test phase suggest that instances of base-rate neglect are driven by stronger attention to common, rather than rare cues. Moreover, the presence of a similarity-by-response interaction is consistent with the inferential prediction that rare and common choices are supported by two or more distinct cognitive mechanisms, including a controlled process that involves reasoning about the well-known category to eliminate it as an option in the face of novel stimuli.

Rare responding on ambiguous trials during test was associated with slowed reaction times compared with common responding. With this finding in mind, we sought to test whether RTs on such trials covaried with subjects' neural similarity to either cue. To examine whether response-varying similarity to cues was predictive of RT differences, we conducted a multilevel model using ln(RT) as the outcome variable and level-1 response, similarity to the chosen cue, and similarity to the unchosen cue as predictor variables within the 22 subjects over 462 observations. Here, similarities to each cue were grouped by whether the cue was consistent or inconsistent with the given response. In the context of a common response, similarity to chosen represented similarity to the common cue, and similarity to unchosen represented similarity to the rare cue; the opposite was true on trials where the rare category was selected. Log-transformed RT was used rather than the raw values to account for the strong positive skew common to RT distributions. Linear modeling including random intercepts for subjects revealed an interaction effect between response (common/rare) and similarity

to the opposite cue from the unchosen category, $\gamma = 0.075$, $t = 2.09$, $p = .037$. Similarity to the cue associated with the chosen response did not vary across responses. Marginal main effects for response type ($\gamma = -0.085$, $t = -2.94$, $p = .003$) and similarity to the unchosen category ($\gamma = -0.070$, $t = -2.75$, $p = .006$) were significant. The interpretation of the positive interaction coefficient as it relates to RT is that when individuals make rare responses, greater neural similarity to the common item is predictive of faster RTs. This pattern is consistent with the inferential hypothesis, as it suggests that selecting the rare category is facilitated by stronger attention to the common cue.

3.5. Neural Correlates of Novelty During the Test Phase

Different theoretical models attempting to explain the IBRE generally assume different quantitative patterns and psychological responses when participants are presented with novel cue combinations in test items. Our task provided an opportunity to contrast mean BOLD activation between pretrained stimulus pairs (*I.PC*, *I.PR*) and novel pairings that were matched for visual features (*I.PCo*, *I.PRo*). The contrast for Novel > Old item pairings revealed a pattern of activation often described as the fronto-parietal control network (Gao and Lin, 2012; Spreng et al., 2013; Liu et al., 2015), including anterior cingulate cortex, dorsolateral PFC, and superior parietal cortex (Figure 6; Table 3).

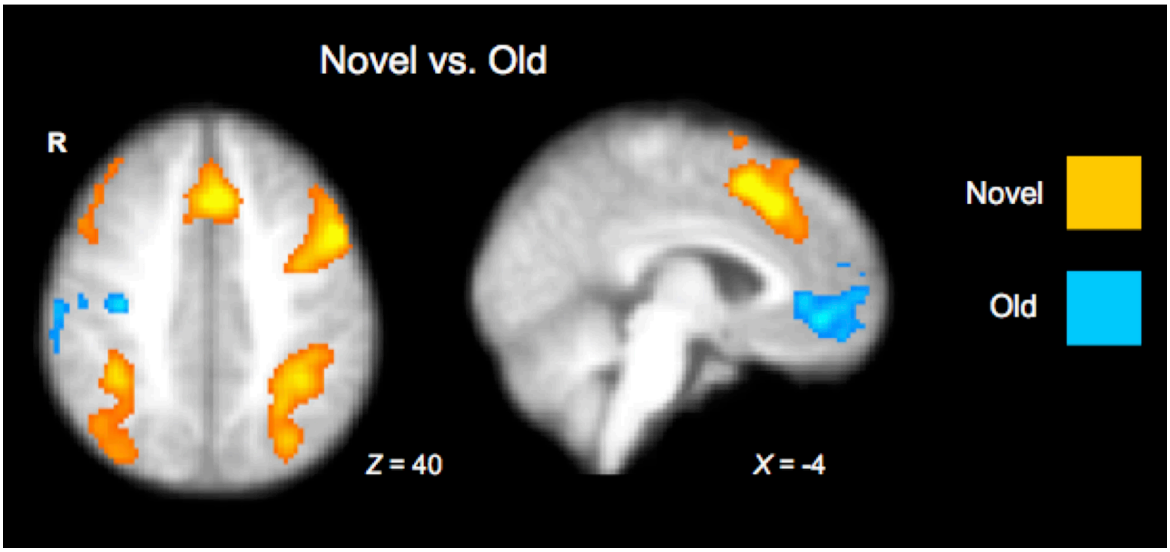

Figure 6. Univariate fMRI results contrasting BOLD activation during the test phase for novel stimulus pairs (*I.PCo* and *I.PRo*) with pairs that had been previously encountered during learning (*I.PC* and *I.PR*).

Consistent with studies on face, object, and scene representation, novel pairings of items were associated with greater mean activation in areas of the ventral occipitotemporal cortex known to represent such categories (Grill-Spector and Weiner, 2014). Contrasting Old > Novel item pairings revealed activation in the vmPFC, a region associated with the application of learned category rules (Liu et al., 2015), and high relative response confidence during decision making (Schnyer, Nicholls, and Verfaille, 2005; Lebreton et al., 2015). The increased activation of lateral and dorsal PFC for the novel cue combinations and vmPFC for previously-learned cue combinations is consistent with recent results suggesting that these regions track decision evidence during categorization tasks (DeGutis and D'Esposito, 2007; Seger et al., 2015; Davis et al., 2017).

Table 3. Activated clusters and peaks (MNI coordinates) for the fMRI results in Figure 6.

Contrast	Cluster Size	Region	*X*	*Y*	*Z*	Max *t*-stat
Novel > Old	Cluster 1 Voxels: 5131	Dorsomedial PFC / Paracingulate Gyrus	0	18	48	7.74
		Left dlPFC / Middle Frontal Gyrus	-52	12	36	7.11
		Left Precentral Gyrus	-38	2	36	5.31
	Cluster 2 Voxels: 1212	Right dlPFC / Middle Frontal Gyrus	48	12	46	4.55
		Right Superior Frontal Gyrus	30	2	64	4.33
	Cluster 3 Voxels: 1610	Right Superior Parietal Lobule	36	-48	42	5.41
		Right Superior Lateral Occipital Cortex	32	-74	48	4.36
	Cluster 4 Voxels: 2060	Left Supramarginal Gyrus	-44	-48	56	6.50
		Left Superior Parietal Lobule	-30	-56	44	5.42
		Left Superior Lateral Occipital Cortex	-30	-72	40	4.71
Old > Novel	Cluster 1 Voxels: 1550	Frontal Pole	-10	62	0	4.17
		vmPFC / Paracingulate Gyrus	2	44	-4	4.35
		vmPFC / Rostral Anterior Cingulate	0	40	2	4.65
	Cluster 2 Voxels: 1223	Left Central Opercular Cortex	-60	-6	8	4.80
		Left Postcentral Gyrus	-64	-22	18	4.16
		Left Parietal Operculum Cortex	-34	-30	18	3.89
		Left Planum Temporale	-40	-34	12	3.61
	Cluster 3 Voxels: 2662	Right Central Opercular Cortex	60	-8	8	5.38
		Right Precentral Gyrus	28	-22	72	6.13

4. Discussion

In this study, we tested two competing hypotheses for the IBRE using fMRI and multi-voxel pattern analysis. Associative theories predict that base-rate neglect is driven by learned selective attention to rare cues. Inferential theories predict that when presented with ambiguous stimuli, people choose rare categories because such items conflict with their well-established knowledge of the common category. To resolve these theories, we used MVPA to decode which cues participants are processing when they make category decisions in the face of conflicting evidence. Our neural similarity measure proved effective at detecting attention to different visual cues during both learning and test phases. In the primary analysis, we found stronger activation of multi-voxel patterns associated with common cues on trials where participants went against disease base-rates by categorizing an ambiguous combination of cues as a member of the rarer category (Figure 5). These results are consistent with the predictions of inferential theories, which suggest that the IBRE arises when participants use their knowledge about the common category to eliminate it as a viable response option (Juslin et al., 2001; Winman et al., 2005). Individuals with stronger activation of patterns associated with the common cue during the learning phase engaged in base-rate neglect more frequently during test than those with weaker representations, further suggesting that it is attention to the regularities of the common category, as opposed to the distinctiveness of the rare category, that leads participants to make rare choices when faced with ambiguous evidence.

The IBRE exemplifies a case in cognitive neuroscience where competing models that predict essentially the same behavioral patterns make very different assumptions

about the cognitive processes, and accordingly, brain states, involved in producing the behavior. Our findings from the test phase (Figure 5) represent a critical step forward in an emerging area of research using multivariate fMRI to reveal that qualitatively distinct brain states may reflect the use of multiple response strategies in the face of identical stimuli (Mack et al., 2013). Consistent with past research using MVPA to decode learned selective attention (Mack et al., 2013; 2016), multi-voxel patterns associated with predictive cues were more strongly activated than imperfectly predictive cues during the learning phase. Using the same approach to decode which information participants were focusing on during ambiguous test trials, we found stronger activation of patterns associated with the common cue (PC) compared to the rare cue (PR). These results are consistent with the inferential hypothesis that activating knowledge of the common category encourages rare responding. No differences in cue activation were observed on trials that received a common response, suggesting that only rare responses depend on deliberative processes: inferential accounts of the IBRE assume that when participants respond according to the base rates, their responses are guided by basic feature-based similarity (Winman et al., 2005).

In line with recent evidence, we identify representational changes indicative of a selective attention mechanism that tracks the predictive value of different stimulus dimensions during learning (Mack et al., 2016; Leong et al., 2017). The data suggest that attention to these goal-relevant cues enhances their representation in object-selective cortex quickly and persistently in a simple rule-based category learning task (Figure 3). However, our findings build upon these results in a crucial way by demonstrating that the attention allocated to competing cues is not static, but rather

dynamically changes based on participants' cognitive inferences and their behavior. Until now, many of the fine-grained dynamics governing how participants choose and switch between strategies are lost because they are unidentifiable in behavior. For example, it is impossible to discern based on behavior alone which hypotheses participants are testing on individual trials before they respond in a consistent manner. Our results suggest that it may be possible to use MVPA methods to synchronize neural and behavioral processes to pinpoint the inferences subjects are considering at a given time.

The present experiment adds to the broader literature on the IBRE by demonstrating the importance of examining processing differences that are dependent on subjects' responses. In the behavioral literature it is common to focus on the average distributions of responses, such as average proportion of rare choices, as opposed to examining differences between choice options for the same stimuli. Here, we show that the way people approach ambiguous test stimuli is not consistent across trials, as individuals appear to engage qualitatively distinct neural processes when producing rare versus common responses. The observed pattern of RT in this experiment reinforces the idea that qualitatively distinct mechanisms may support common versus rare choices, as we found a significant slowing of RTs on ambiguous trials when subjects chose the rare disease. This result suggests that choosing the rare response places more cognitive demand on participants, as predicted by our inferential account. While RT differences for rare versus common responses have not been examined in previous studies of the IBRE, we expect that the RT distributions of existing data would reveal response-dependent patterns similar to those found here.

Future research will benefit by incorporating such results into computational models.

Contrasting mean BOLD activation on test trials composed of previously unseen *I.PCo*/*I.PRo* stimulus combinations with previously learned pairs revealed activation in a fronto-parietal network associated with rule-based category learning (Filoteo et al., 2005; Seger and Cincotta, 2006, Soto et al., 2013) and decisional uncertainty (DeGuits and D'Esposito, 2007; Seger et al., 2015; Davis et al., 2017). This novelty-sensitive network included clusters in dorsomedial PFC, dorsolateral PFC, and superior parietal lobule. Alternatively, exposure to previously learned pairs at test was accompanied by activation of vmPFC, which has been associated with attention to strong predictors (Sharpe and Killcross, 2015; Nasser et al., 2017), high confidence / high relative decision evidence (Schnyer, Nicholls, and Verfaille, 2005; Lebreton et al., 2015; Davis et al., 2017) and the application of familiar rules (Boettiger and D'Esposito, 2005). Thus the current mean BOLD activation results are in accordance with a number of recent studies examining decision-related processes in categorization. Moreover, they are consistent with the idea that novel pairings may involve more active deliberation whereas the previously learned pairs do not.

One critical question is whether the trade-off between lateral/dorsal PFC activation for novel stimuli and vmPFC activation for previously learned stimuli reflects wholly separate processes or opposite ends of a common on-task/off-task network. Consistent with the idea that these activation patterns may indicate distinct processes, the regions activated in response to pair novelty mirror those associated with model-based learning in the reinforcement learning literature (Gläscher et al., 2010; Niv et al., 2015), which involves forming high-level representations of the contingencies between

stimuli and outcomes (Daw, Niv, and Dayan, 2005). Likewise, more habitual, model-free learning has been associated with the presence of prediction error signals in ventral striatum (Gläscher et al., 2010; Niv et al., 2015). The prefrontal regions activated for the previously learned stimuli in our study are part of a dopaminergic corticostriatal loop (Seger, 2008), and thus may reflect the maintenance of previously reinforced behaviors in the absence of feedback (Noonan et al., 2012).

The trade-off between model-based and model-free choice may also be useful for understanding the stimulus- and response-dependent neural activation patterns found in the present study at a mechanistic level. For example, recent findings in associative learning have shown that participants who tend to engage in rule-based generalization for ambiguous cue pairings in a patterning task (Shanks and Darby, 1998) tend to rely more on model-based choice (Don et al., 2016). Within individuals, model-based inference may be favored in cases of high uncertainty, while competing for control over responding with the habitual system from trial to trial (Daw, Niv, and Dayan, 2005). A similar dynamic may be involved in producing the stimulus- and response-dependent neural activation patterns found in the present study: when viewing test stimuli, participants might be prompted to use explicit rules only when a mismatch is detected between two competing cue-response mappings. Conversely, treating the pair of cues as a unitary stimulus may allow respondents to rely on efficient, well-established response patterns that reliably favor the common category. A question for future research is whether models that allow for flexible system recruitment (e.g., Ashby et al., 1998; Dayan, 2007) could account for both the behavioral and neural dynamics of the IBRE.

Although a trade-off between anticorrelated neural systems may have contributed to the present findings, another possibility is that our neural data reflect the operation of distinct, but complimentary attentional mechanisms. It has been recently discussed that two such mechanisms may coexist in associative learning: automatic attention akin to that proposed by Mackintosh (1975) that converges on strong predictors, and another concerned with the controlled selection of stimuli that are unreliable (Pearce and Mackintosh, 2010; Esber and Haselgrove, 2011; Luque et al., 2016). Associative attention learning models rely on the formalisms of Mackintosh (1975), where the role of attention is to selectively process cues with the highest predictive value. Alternatively, Pearce and Hall (1980) proposed that attention instead allocates resources to stimuli that have been associated with the most uncertainty, promoting learning and error reduction on future trials. Accordingly, one possible explanation for our multivariate results for the key ambiguous (*PC.PR*) cue pairs is that activation of the common cue is attributable to uncertainty-related attention. In an event-related potential (ERP) study of the IBRE, Wills and colleagues (2014) found evidence to suggest enhanced attention to cue *PR* relative to *PC* following training. However, this evidence came from an index of early visual processing, where the present analysis examines the stimulus presentation period as a whole. It is therefore possible that a rapid form of attention focusing on predictive utility was present during our experiment, with this effect being washed out by the subsequent effects of controlled attention on the ambiguous trials. Consistent with this idea, we found medial prefrontal regions associated with attention to strong predictors (Sharpe and Killcross, 2015; Nasser et al., 2017) to be engaged when participants encountered previously-

trained cue pairings relative to novel pairings during the test phase. If both attentional processes play a role in base-rate neglect, future studies combining ERP and fMRI might help to dissociate these separate signals.

A critical question for future research is how changes in cognitive state from dual task interference may influence the IBRE. Because our inferential account is assumed to depend on deliberative processes such as representing rule information in working memory, it follows that requiring participants to engage in a concurrent task that demands WM resources should diminish their ability to apply such rules, resulting in fewer rare responses on ambiguous *PC.PR* trials. Although Lamberts and Kent (2007) showed that neither the inclusion of a secondary task nor forcing speeded responses during test disrupted the IBRE, the effect requires specific, asymmetrical learning conditions that suggest its occurrence is unlikely to be a product of test strategy alone (Don and Livesey, 2016). Rather, experiments in category learning (Waldron and Ashby, 2001; Zeithamova and Maddox, 2006) and predictive learning (Wills et al., 2011) suggest that adding cognitive load during the learning phase may disrupt rule acquisition and result in more habitual, feature-based generalization patterns at test.

The present findings speak to a larger debate surrounding whether human associative learning is necessarily tied to reasoning (Mitchell et al., 2009). We do not take the apparent involvement of higher-level processes in a behavior that can be captured by associative models to mean that humans are purely propositional beings. Rather, the results at hand lend support to the general view that human learning involves both propositional reasoning and some fundamental associative learning processes, while illustrating that neural representations of the same physical stimulus

can differ substantially depending on which response (or strategy) is selected at a given time.

4.1. Conclusions

While associative theories of base-rate neglect suggest that simple error-reduction mechanisms lead individuals to form an especially strong association between rare cues and their outcomes, here we provide neurobiological evidence that such processes alone are unlikely to account for the generalization patterns that produce the IBRE. Rather, in a novel application of fMRI and multi-voxel pattern analysis, we demonstrate that when selecting a rarer category in the face of ambiguous evidence, people tend to focus on the cues they have encountered the most frequently. Taken together, these findings strongly suggest that high-level processes contribute to the IBRE. Accordingly, computational models that posit a purely associative explanation for base-rate neglect will require modification to account for the pattern of neural data obtained in this experiment. Using multi-voxel pattern analysis to measure the strength of activation for co-present (or unpresent) visual cues in occipitotemporal cortex appears to be a promising avenue to test theories of learned selective attention moving forward.

AUTHOR CONTRIBUTIONS

SO, TD, EL, and DW designed the study and wrote the paper. SO and TD analyzed the fMRI data. TD ran the fMRI participants and behavioral pilots for the experiment.

ACKNOWLEDGEMENTS

Data collection for this project was funded by start-up funds to TD. Funding for preliminary data analysis and collaboration was provided via Visiting International Scholar award from the University of Sydney to TD and EL.

REFERENCES

Abdulrahman H, Henson RN (2016) Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. *NeuroImage* 125: 756-766.

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.

Boettiger CA, D'Esposito M (2005) Frontal networks for learning and executing arbitrary stimulus-response associations. *J Neurosci* 25: 2723–2732.

Bravata DM (2000). Making medical decisions under uncertainty. *Semin Med Pract* 3: 6-13.

Casscells W, Schoenberger A, Graboys TB (1978) Interpretation by physicians of clinical laboratory results*. New Engl J Med* 299: 999.

Davis T, Poldrack RA (2013) Measuring neural representations with fMRI: practices and pitfalls. *Ann NY Acad Sci* 1296: 108-134.

Davis T, LaRocque KL, Mumford JA, Norman KA, Wagner AD, Poldrack RA (2014) What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage* 97: 271-283

Davis T, Poldrack RA (2014) Quantifying the internal structure of categories using a neural typicality measure. *Cereb Cortex* 24: 1720–1737.

Davis T, Xue G, Love BC, Preston AR, Poldrack RA (2014) Global neural pattern similarity as a common basis for categorization and recognition memory. *J Neurosci* 34: 7472-7484.

Davis T, Goldwater MB, Giron J (2017) From concrete examples to abstract relations: The rostrolateral prefrontal cortex integrates novel examples into relational categories. *Cereb Cortex* 27: 2652-2670.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8: 1704–1711.

Dayan P (2007) Bilinearity rules and prefrontal cortex. *Front Comp Neurosci* 1: 1-14.

DeGutis J, D'Esposito M (2007). Distinct mechanisms in visual category learning. *Cogn Affect Behave Neurosci* 7: 251-259.

Don H, Goldwater M, Otto R, Livesey E (2016) Rule abstraction, model-based choice, and cognitive reflection. *Psychon Bull Rev* 23: 1615-1623.

Don H.J, Livesey EJ (2016) Effects of outcome and trial frequency on the inverse base-rate effect. *Mem Cogn*, in press.

Esber GR, Haselgrove M (2011) Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proc Biol Sci* 278: 2553-2561.

Filoteo JV, Maddox WT, Simmons AN, Ing AD, Cagigas XE, Matthews S, Paulas MP (2005) Cortical and subcortical brain regions involved in rule-based category learning. *Neuroreport* 16: 111-115.

Gao W, Lin W (2012) Frontal parietal control network regulates the anti-correlated default and dorsal attention networks. *Hum Brain Mapp* 33: 192-202.

Gazzaley A, Nobre AC (2011). Top-down modulation: bridging selective attention and working memory. *Trends Cogn Sci* 16: 121-139.

Gläscher J, Daw N, Dayan P, O'Doherty J (2010) States versus rewards: Dissociable prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66: 585-595.

Grill-Spector K, Weiner KS (2014) The functional architechture of ventral temporal cortex and its role in categorization. *Nat Rev Neurosci* 15: 536-548.

Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby JV, Pollmann S (2009) PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics* 7: 37-53.

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: 2425–2430.

Haxby, JV, Connolly, AC, Guntupalli, JS (2014) Decoding neural representational spaces using multivariate pattern analysis. *Ann Rev Neurosci* 37: 435-456.

Hollerman JR, Shultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci* 1: 304-309.

Ishai, A, Ungerleider LG, Martin A, Schouten JL, Haxby JV (1999) Distributed representation of objects in the human ventral visual pathway. *Proc Natl Acad Sci USA* 96: 9379.

Johansen MK, Fouquet N, Shanks DR (2007) Paradoxical effects of base rates and representation in category learning. *Mem Cogn* 35: 1365-1379.

Juslin P, Wennerholm P, Winman A (2001) High-level reasoning and base-rate use: Do we need cue-competition to explain the inverse base-rate effect? *J Exp Psychol Learn Mem Cogn* 27: 849-871.

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: 4.

Kruschke JK (1996) Base rates in category learning. *J Exp Psychol Learn Mem Cogn* 1: 3-26.

Kruschke JK (2001a) The inverse base-rate effect is not explained by eliminative inference. *J Exp Psychol Learn Mem Cogn* 27: 1385-1400.

Kruschke JK (2001b) Toward a unified model of attention in associative learning. *J Math Psychol* 45: 812-863.

Kruschke JK (2003) Attentional theory is a viable explanation of the inverse base-rate effect: A reply to Winman, Wennerholm, and Juslin (2003). *J Exp Psychol Learn Mem Cogn* 29: 1396-1400.

Lamberts K, Kent C (2007) No evidence for rule-based processing in the inverse base rate effect. *Mem Cognit* 35: 2097–2105.

Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* 18: 1159-1170.

Leong YC, Radelescu, A, Daniel, R, Niv, Y (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron* 93: 451-463.

Lewis-Peacock JA, Postle BR (2008) Temporary activation of long-term memory supports working memory. *J Neurosci* 28: 8765–8771.

Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR (2012) Neural evidence for a distinction between short-term memory and the focus of attention. *J Cogn Neurosci* 24: 61-79.

Liu Z, Braunlich K, Wehe HS, Seger CA (2015) Neural networks supporting switching, hypothesis testing, and rule application. *Neuropsychologia* 77: 19-34.

Luque D, Vadillo MA, Le Pelley ME, Beesley T (2016) Prediction and uncertainty in as sociative learning: examining controlled and automatic components of learned attentional biases. *Q J Exp Psychol*, published online June 7, 2016, doi: 10.1080/17470218.2016.1188407

Mack ML, Preston AR, Love BC (2013) Decoding the brain's algorithm for categorization from its neural implementation. *Curr Bio* 23: 1-5.

Mack ML, Love BC, Preston AR (2016) Dynamic updating of hippocampal representations reflects new conceptual knowledge. *Proc Natl Acad Sci USA* 113: 13203- 13208.

Mackintosh NJ (1975) A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychol Rev* 82: 276-298.

Mitchell CJ, De Houwer J, Lovibond P (2009) The propositional nature of human associative learning. *Behav Brain Sci* 32: 183–198.

Medin DL, Edelson SM (1988) Problem structure and the use of base-rate information from experience. *J Exp Psychol Gen* 117: 68-85.

Medin DL, Bettger JG (1991) Sensitivity to changes in base-rate information. *Am J Psychol* 104: 311–332.

Milton FN, Bealing P, Carpenter KL, Benattayallah A, Wills AJ (2016) The neural correlates of similarity- and rule-based generalization. *J Cognitive Neurosci* 29: 150-166.

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.

Mumford JA, Davis T, Poldrack RA (2014) The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage* 103: 130-138.

Nasser HM, Calu DJ, Schoenbaum G, Sharpe MJ (2017) The dopamine prediction error: contributions to associative models of reward learning. *Front Psychol* 8: 244.

Niv Y, Daniel R, Deana J, Gershman SA, Leong YC, Radulescu A, Wilson RC (2015) Reinforcement learning in multidimensional environments relies on attention mechanisms. *J Neurosci* 35: 8145-8157.

Nomura EM, Maddox WT, Filoteo JV, Ing AD, Gitelman DR, Parrish TB, Reber PJ (2007) Neural correlates of rule-based and information-integration visual category learning. *Cereb Cortex* 17: 37–43.

Noonan MP, Kolling N, Walton ME, Rushworth MFS (2012) Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement. *Eur J Neurosci* 35: 997-1010.

Pearce JM, Hall G (1980) A model for pavlovian learning—variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol Rev* 87: 532–552.

Pearce JM, Mackintosh NJ (2010) Two theories of attention: a review and possible integration. In *Attention and Associative Learning, From Brain to Behaviour*, eds C. Mitchell and M. E. Le Pelley (Oxford: Oxford University Press), pp. 11–40.

Rissman J, Gazzaley A, D'Esposito M (2004) Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage* 23: 752–763.

Rissman, J, Wagner, AD (2012) Distributed representations in memory: insights from functional brain imaging. *Ann Rev Psychol* 63: 101-128.

Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL, Petersen SE (2014) Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Hum Brain Mapp* 35: 1981–1996.

Schlicting ML, Mumford JA, Preston AR (2015) Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat Commun* 6: 8151.

Schnyer DM, Nicholls L, Verfaellie M (2005) The role of VMPC in metamemorial judgments of content retrievability. *J Cognitive Neurosci* 17: 832-846.

Schnyer DM, Maddox WT, Ell S, Davis S, Pacheco J, Verfaellie M (2009) Prefrontal contributions to rule-based and informa- tion-integration category learning. *Neuropsychologia* 47: 2995–3006.

Seger CA, Cincotta CM (2006) Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cereb Cortex* 16: 1546–1555.

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, Miller EK (2010) Category learning in the brain. *Annu Rev Neurosci* 33: 203–219.

Seger CA, Braunlich K, Wehe HS, Liu Z (2015) Generalization in category learning: The roles of representational and decisional uncertainty. *J Neurosci* 35: 8802-8812.

Shanks DR, Darby RJ (1998) Feature- and rule-based generalization in human associative learning. *J Exp Psychol Anim Behav Process* 24: 405-415.

Sharpe MJ, Killcross S (2015) The prelimbic cortex directs attention toward predictive cues during fear learning. *Learn Mem* 22: 289-293.

Smith EE, Grossman M (2008) Multiple systems of category learning. *Neurosci Biobehav Rev* 32: 249–264.

Soto FA, Waldschmidt JG, Helie S, Ashby FG (2013) Brain activity across the development of automatic categorization: A comparison of categorization tasks using multi-voxel pattern analysis. *NeuroImage* 71: 284-297.

Spreng RN, Sepulcre J, Turner GR, Stevens DW, Schacter DL (2013) Intrinsic architecture underlying the relations among the default, dorsal attention, and frontoparietal control networks of the human brain. *J Cognitive Neurosci* 25: 74-86.

Turner BO, Mumford JA, Poldrack RA, Ashby GF (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage* 62: 1429-1438.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Sciences* 185: 1124-1131.

Waldron EM, Ashby FG (2001) The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychon Bull Rev* 8: 168–176.

Wills AJ, Graham S, Koh ZS, McLaren IPL, Rolland MD (2011) Effects of concurrent load on feature- and rule-based generalization in human contingency learning. *J Exp Psychol Anim Behav Process* 37: 308–316.

Wills AJ, Lavric A, Hemmings Y, Surrey, E (2014) Attention, predictive learning, and the inverse base-rate effect: Evidence from event-related potentials. *NeuroImage* 87: 61-71.

Wimmer, GE, Shohamy, D (2012) Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* 338: 270-273.

Winman A, Wennerholm P, Juslin P (2003) Can attentional theory explain the inverse base-rate effect? Comments on Kruschke (2001). *J Exp Psychol Learn Mem Cogn* 29: 1390-1395.

Winman A, Wennerholm P, Juslin P, Shanks, DR (2005) Evidence for rule-based processes in the inverse base-rate effect. *Q J Exp Psychol* 58A: 789-815.

Zeithamova D, Maddox WT (2006) Dual-task interference in perceptual category learning. *Mem Cogn* 34: 387– 398.

Zeithamova D, Dominick AL, Preston, AR (2012) Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* 75: 168–179.

SUPPLEMENTARY MATERIAL

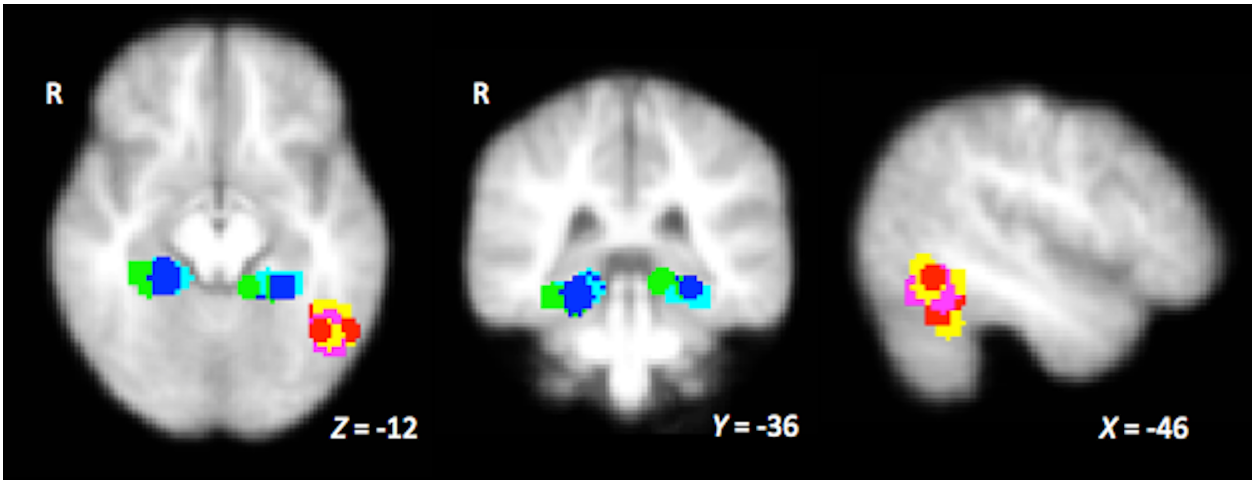

Figure A1. Distribution of object- and scene-selective ROIs across subjects for the RSA test analysis. Each 6-mm sphere surrounds subjects' peak activation to each respective visual stimulus category within anatomically defined regions (objects: left inferior posterior temporal gyrus; scenes: bilateral parahippo-campal gyrus). Object-selective ROIs are depicted in warm colors; scene-selective ROIs are depicted in cool colors.

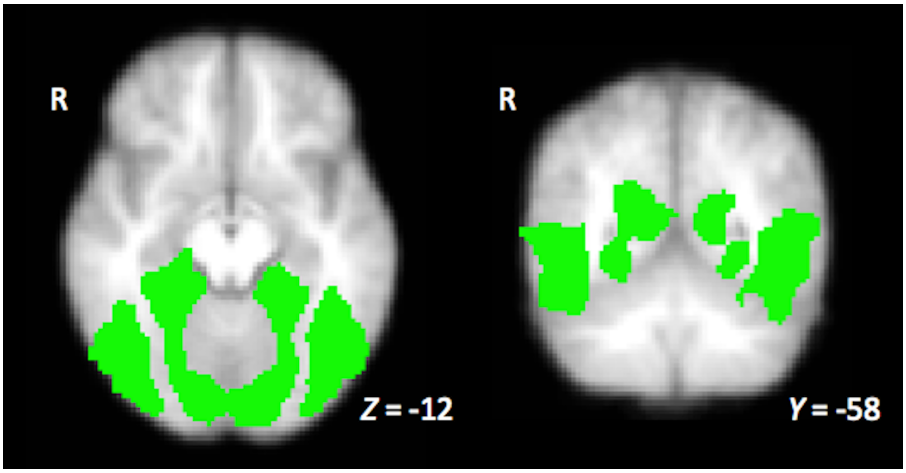

Figure A2. Functional localizer mask encompassing face-, object-, and scene-selective regions for the RSA learning analysis.

Table A1. Peaks (MNI coordinates) for the subject-by-subject object- and scene-selective ROIs depicted in Figure A1.

	Objects > Scenes				**Scenes > Objects**		
Subject Number	*X*	*Y*	*Z*		*X*	*Y*	*Z*
1	-48	-60	-8		18	-36	-8
2	-48	-56	-12		30	-38	-12
3	-42	-50	-16		24	-38	-14
4	-42	-56	-8		28	-34	-12
5	-50	-60	-8		-14	-40	-6
6	-50	-52	-10		24	-32	-18
7	-58	-50	-26		-30	-38	-14
8	-52	-60	-10		-22	-40	-12
9	-54	-46	-22		18	-36	-10
10	-42	-58	-8		18	-36	-12
11	-42	-62	-4		34	-34	-14
12	-46	-52	-26		-16	-40	-8
13	-48	-58	-14		28	-30	-20
14	-44	-62	-10		24	-36	-16
15	-44	-50	-10		24	-36	-12
16	-44	-56	-10		-20	-38	-12
17	-42	-62	-4		-24	-40	-10
18	-54	-58	-8		24	-32	-16
19	-48	-56	-22		-12	-40	-8
20	-46	-54	-16		-14	-40	-6
21	-48	-64	-14		22	-36	-14
22	-46	-60	-12		-26	-40	-10