Cognitive, Affective, & Behavioral Neuroscience
in press

The impact of training methodology and representation on rule–based categorization: An fMRI study

Sébastien Hélie, Farzin Shamloo
Department of Psychological Sciences, Purdue University

Hanru Zhang
Department of Psychology, Colorado State University

Shawn W. Ell
Department of Psychology, University of Maine

Hélie, Shamloo, & Ell (2017) showed that regular classification learning instructions (A/B) promote between–category knowledge in rule–based categorization whereas conceptual learning instructions (YES/NO) promote learning within–category knowledge with the same categories. Here we explore how these tasks affect brain activity using fMRI. Participants learned two sets of two categories. Computational models were fit to the behavioral data to determine the type of knowledge learned by each participant. fMRI contrasts were computed to compare BOLD signal between the tasks and between the types of knowledge. The results show that participants in the YES/NO task had more activity in the pre–supplementary motor area, prefrontal cortex, and the angular/supramarginal gyrus. These brain areas are related to working memory and part of the dorsal attention network, which showed increased task–based functional connectivity with the medial temporal lobes. In contrast, participants in the A/B task had more activity in the thalamus and caudate. These results suggest that participants in the YES/NO task used bivalent rules and may have treated each contextual question as a separate task, switching task each time the question changed. Activity in the A/B condition was more consistent with participants applying direct Stimulus → Response rules. With regards to knowledge representation, there was a large shared network of brain areas, but participants learning between–category information showed additional posterior parietal activity, which may be related to the inhibition of incorrect motor programs.

Keywords: category representation, rule–based categorization, fMRI

Introduction

Categorical representations are the building blocks of decision–making from the most routine to the most novel contexts (Hélie & Ashby, 2012). Not surprisingly, the study of the processes underlying the development of the representations necessary for such decision making has been the focus of much research (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Maddox & Ashby, 1993; Markman & Ross, 2003; J. D. Smith & Minda, 2002). It has been argued that category representations can be broadly classified as containing within–category information (what is common to category members, e.g., humans generally have one head) or between–category information (what is different between members of different categories, e.g., dogs generally have more legs than humans). The above examples show an important difference between within– and between–category representations. Within–category representations directly describe the category members (e.g., how many heads humans have) without any reference to other categories (e.g., it says nothing about dogs). In contrast, between–category representations do not directly qualify category members. They compare members from one category to members of other categories. Here, dogs have more legs than humans, but the rule does not say how many legs dogs (or humans) have. Hélie, Ell, and colleagues showed that participants typically learn within–category information in concept training (YES/NO) and inference learning tasks (Ell, Smith, Peralta, & Hélie, 2017; Ell, Smith, Deng, & Hélie, 2020), but that classification training (A/B) with rule–based categories biases participants towards learning between–category information (Hélie, Shamloo, & Ell, 2017).

The goal of this article is to explore the brain circuits underlying the learning of these two tasks (YES/NO, A/B) and types of representations (within–category, between–category). Towards this goal, we replicated Experiment 1 from Hélie, Shamloo, and Ell (2017) using a functional Magnetic Resonance Imaging (fMRI) rapid event–related design. Specifically, participants learned two sets of two categories during training and were subsequently tested on a novel categorization problem using the training categories. The stimuli were sine–wave gratings varying in bar width (frequency) and orientation (counterclockwise rotation from horizontal). The categories are shown in Figure 1 and each symbol rep-

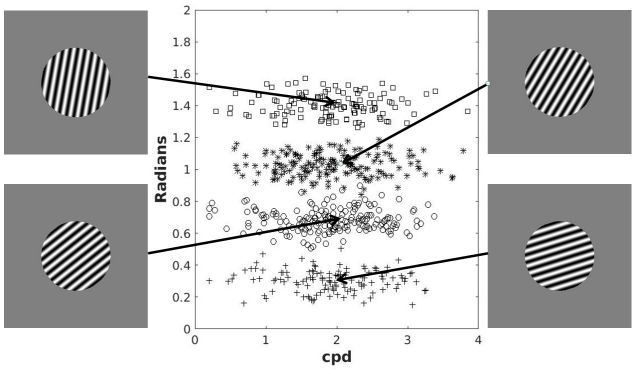

Figure 1. Stimuli used in the Experiment. The *x*-axis corresponds to the width of the bars (frequency) and the *y*-axis corresponds to the rotation angle of the bars (counterclockwise from horizontal). Symbols denote different categories. The mean stimulus of each category is shown as an example.

resents the coordinates used to draw one stimulus (i.e., one specific frequency and rotation angle). Different symbols are used to represent the four categories. During training, participants learned to distinguish stimuli in the '+' category (arbitrary labelled "A") from those in the '∘' category (arbitrary labelled "B"). They also learned to distinguish stimuli in the '*' category (arbitrary labelled "C") from those in the '□' category (arbitrary labelled "D"). At test, participants were asked to distinguish between stimuli in the "B" and "C" categories (shown as '∘' and '*' in the figure). Critically, participants were already familiar with the test categories but had never been asked to contrast these two categories. If participants learned within–category representations during training, then this knowledge should seemlessly transfer to the test phase (because participants are already familiar with the categories). However, if the participants learned between–category representations during training, then they should not be able to apply their knowledge in the test phase (because the test comparison is new).

Hélie, Shamloo, and Ell (2017) trained participants in the above task using one of two training methodologies. In A/B training, participants were asked in each training trial a specific categorization question, e.g., "A or B?". In this case, participants knew that the stimulus on the screen was either a member of the "A" or "B" category and needed to push a button corresponding to the correct category label. Other possible questions were "C or D?" (also at training) and "B or C?" (at test). In the YES/NO training condition, participants saw a different type of categorization questions. An example question might be "Is this an A?". The participant's task was to press either the "yes" or "no" button depending on how they categorized the stimulus on the screen. Similar questions were presented for categories "B", "C", and "D". The results in Hélie, Shamloo, and Ell (2017) showed that A/B training resulted in a bias toward between–category representations (with high transfer accuracy cost at test) whereas YES/NO training resulted in a bias toward within–category representations (with low transfer accuracy cost at test).

Hypotheses

Hélie, Roeder, and Ashby (2010) showed that rule–based learning with similar sine–wave gratings involves a brain circuit centered around the ventrolateral prefrontal cortex (PFC) partly relying on working memory (WM) (Ashby, Ell, Valentin, & Casale, 2005). However, the data from Hélie et al. were collected using an A/B paradigm only, and no attempt was made to identify the type of category representations learned. As a result, they likely included both participants learning rules containing within–category information and participants learning rules containing between–category information. In the current study, some participants were trained using a A/B task while others were trained using a YES/NO task. In addition to comparing brain activity between these two categorization tasks, computational models were fit to the behavioral data to identify the type of category representations learned by each participant (Hélie, Shamloo, & Ell, 2017). The fit of the computational models to each individual participant were used as weights to explore blood oxygen–level dependent (BOLD) response of participants learning within– or between–category representation, regardless of training task. We hypothesized that the difference between the types of learning task (A/B vs. YES/NO) and rule content (between– vs. within–category information) would show different circuits reflecting either the task demands or knowledge content.

Main task–based hypotheses. With regards to tasks, Hélie et al. (2010) found that when learning rules in an A/B task with multiple decision bounds (as used in the present experiment), categorization accuracy was linked to brain acitivty in the medial temporal lobes (MTL), anterior cingulate cortex, and thalamus. Several clusters of BOLD acitivty were also found in the PFC. These rules are univalent: the same stimulus should always produce the same button press (Bunge, 2004). We expect to reproduce this result in the A/B condition. However, because the YES/NO task does not include a consistent Stimulus → Response mapping, we expect the rules learned in this task to be bivalent, meaning that a given stimulus requires a different button press depending on context. Hence, the former should be represented more rostrally in the PFC (Badre, Kayser, & D'Esposito, 2010). Also, after the stimulus category has been determined in the YES/NO task one still needs to disentangle the motor responses and decide which button to press. This should produce activity in the pre–supplementary motor area (preSMA) because participants need to "pay attention to their intentions" (Nachev, Kennard, & Husain, 2008, p. 858).

The explanation above suggests another interesting possibility. If participants in the YES/NO task are using the

question as a context for response selection, then it is possible that each question is considered its own task, meaning that only rules related to the question currently on screen are maintained in WM and used to respond (Fleischer & Hélie, 2020; Schneider & Logan, 2014). If this is the case, we expect to detect brain activity generally related to task–switching for this condition, namely the frontal pole (Wang et al., 2010) and the supramarginal gyrus (Philipp, Weidner, Koch, & Fink, 2013).

Exploratory hypotheses: Model–based training. With regards to knowledge representation, Hélie, Ell, Filoteo, and Maddox (2015) argued that processing in the PFC could emulate decision bounds and that categorization rules could be treated as regular WM items (Fleischer & Hélie, 2020). As a result, we predict activity in the PFC, temporal lobe, and posterior parietal cortex for participants learning between–category representations (Ashby et al., 2005). Next, Zeithamova, Maddox, and Schnyer (2008) showed that concept learning produced activity in the superior parietal lobule and inferior lateral occipital cortex (see their Table 2). However, the task in Zeithamova et al. differed from the current task in important ways. Most importantly, the stimuli were cartoon animals composed of 10 binary dimensions. In contrast, the present experiment used sine–wave gratings with two continuous dimensions. We nevertheless expect to reproduce their results and find activity in the superior parietal lobule and inferior lateral occipital cortex for participants learning within–category representations.

Exploratory hypotheses: Task–based training functional connectivity. Hélie et al. (2010) found that initial category learning of rule–based categories was associated with MTL activity. Likewise, Nomura and Reber (2008) also found that rule–based categorization was associated with increased MTL activity. Several other category learning studies not assuming rule–based representations also found activity in the MTL (e.g., Bowman & Zeithamova, 2018; Zeithamova et al., 2019). However, Hélie et al. did not find any cluster of BOLD signal in the MTL: activity in the MTL only reached statistical significance in region of interest analyses with a more relaxed statistical significance threshold (because of a reduced need for mutiple–testing correction). The region of interest analysis in Hélie et al. was inspired by COVIS (Ashby et al., 1998; Ashby & Valentin, 2017), which predicts that the MTL is part of a network linked to hypothesis–testing. The present experiment includes more participants than previous work by Hélie et al., which should increase statistical power and improve detection. However, the stimuli are simple (sine–wave gratings) and the MTL has been shown to play a more prominent role with stimuli that include complex spatial features (Lee, Yeung, & Barense, 2012). This suggests that even with a larger sample size MTL activity may go undetected in the present experiment. For these reasons, we also performed a psychophysiological interaction (PPI) analysis with the MTL as a seed to explore task–related functional connectivity with the MTL. While this analysis is exploratory, we predicted that learning categorization rules would increase functional connectivity with brain areas related to WM, and in particular with the dorsal attention network (Majerus, Péters, Bouffier, Cowan, & Phillips, 2017; Vossel, Geng, & Fink, 2014). In addition, increased functional connectivity with the cognitive control network should be more pronounced in the YES/NO task, as this task requires using context dependent rules.

Exploratory hypotheses: Task–based test. The experiment also included a test phase where accuracy feedback was removed and a new category comparison was introduced (i.e., "B" vs. "C"). Behavioral data from the new category comparison is necessary to fit the models and identify the type of representation learned by the participants (i.e., within–category vs. between–category; see Hélie, Shamloo, & Ell, 2017). However, it is possible that removing the feedback may affect the way participants process the tasks (e.g., a change in strategy) and also affect BOLD response. While we do not expect that participants would process the test block differently, brain activity during the test block was analyzed.

Materials and methods

This experiment reproduced Experiment 1 from Hélie, Shamloo, and Ell (2017) in a MRI scanner.

Participants

Forty–three students from Purdue University were recruited to participate in this experiment (17 males). Twenty participants were randomly assigned to the YES/NO condition while the remaining 23 participants were assigned to the A/B condition. One participant in the A/B condition did not complete the experiment because of claustrophobia. Each participant received $30 as compensation for their time. All procedures were approved by the Purdue University Biomedical Institutional Review Board.

Stimuli and apparatus

The stimuli were circular sine–wave gratings of constant contrast and size backprojected on a mirror attached to a head coil using a Hyperion HD 1080p projector ($1,920 \times 1,080$ resolution). Each stimulus was defined in a 2D space by a set of points $(frequency, orientation)$ where $frequency$ (bar width) was calculated in cycles per degree (cpd), and $orientation$ (counterclockwise rotation from horizontal) was calculated in radians. The stimuli were generated with Matlab using the Psychophysics toolbox (Brainard, 1997) and occupied an approximate visual angle of 5°. In each trial, a single stimulus was presented in the center of the screen.

The categories used for both training conditions are shown in Figure 1. There were four separate categories generated using bivariate normal distributions with the randomization technique (Ashby & Gott, 1988). The categories were arbitrarily labeled with letters A–D from bottom to top. Only the mean orientation differed across categories, so that $\mu_A = (1.9, 0.30)$, $\mu_B = (1.9, 0.67)$, $\mu_C = (1.9, 1.03)$, and $\mu_D = (1.9, 1.40)$. The covariance matrix for all categories was $\Sigma = \left(\begin{smallmatrix} 0.44 & 0 \\ 0 & 0.01 \end{smallmatrix}\right)$. This yielded stimuli that varied in orientation from 10° to 90° (counterclockwise from horizontal) and in bar width (frequency) between 0.2 and 3.85 cpd. Note that the categories lie on a continuum of orientation, and that the width of the bars was irrelevant. Specifically, these categories can be near–perfectly separated by three linear boundaries corresponding to the following verbal rule: near–horizontal stimuli are "A", slightly steeper stimuli are "B", much steeper stimuli are "C", and near vertical stimuli are "D". A single set of 600 stimuli was generated from these distributions, and the stimulus set was linearly transformed so that the sample mean and covariance of each category matched the generative distributions. The resulting set of stimuli is shown in Figure 1. The stimulus set was independently shuffled for each participant.

Stimulus presentation, feedback, and response recording were controlled and acquired using Matlab. Responses were produced by using two MR–compatible Celeritas button boxes (one in each hand, with three buttons on each box). In the A/B condition, the following question was displayed in the top–middle of the screen in black font "X or Y?" where X and Y were replaced by category labels informing the participants that they should respond using one of these two categories in this trial. The left button in the left hand corresponded to an "A" response, the right button in the left hand corresponded to the "B" response, the left button in the right hand corresponded to the "C" response, and the right button in the right hand corresponded to the "D" response. The middle button in each hand was not used.

In the YES/NO condition, the following question was displayed in the top–middle of the screen in black font: "Is this a 'X'?", where X was replaced by a category label informing the participants of the target category for this trial. All buttons in the left hand corresponded to the "YES" response while all buttons in the right hand corresponded to the "NO" response.

In both conditions, visual feedback was given for a correct (a green checkmark) or incorrect (a big red "X") response. If a response was too late, participants saw a big black dot. During the whole experiment, the screen background was gray.

Study design

The experiment was composed of 6 blocks of 100 trials (for a total of 600 trials), and each stimulus was seen only once. Participants were told they were taking part in a categorization experiment and that they had to assign each stimulus into either an "A", "B", "C", or "D" category. The participants were told that there would be a test phase at the end of the experiment where they would no longer be receiving feedback. No further detail was given about the test phase at this point.

The first 5 blocks were training blocks and the participants were trained to separate "A" stimuli from "B" stimuli and "C" stimuli from "D" stimuli. Specifically, only the questions "A or B?" and "C or D?" were used in the A/B condition. Likewise, if the question was "Is this an 'A'?" in the YES/NO condition, correct "NO" responses were from the "B" category, and if the question was "Is this a 'B'?", correct "NO" responses were from the "A" category (and the same logic applied to the "C" and "D" categories). All categories and questions were equally likely. After the training phase, the participants were told that they were now beginning the test phase, and that they should use the categories learned during the training phase to respond in the test phase. Instead of feedback, participants saw a large texture pattern on the screen. The texture was neutral, always the same, and participants were told that the texture was non–informative.[1] Note that the training and test stimulus sets were non–overlapping and randomly selected for each participant.

The rapid–event related design had three types of events: (1) stimulus presentation, (2) feedback presentation, and (3) blank screen. The timing of a trial [scaled in repetition time (TR), 1 TR = 720 ms] went as follows: (1) a response–terminated stimulus was presented for 3 TR. If the participant responded in less than 3 TR, the stimulus disappeared and was replaced by a blank screen for the remainder of the 3 TR. Next, feedback was presented for 1 TR. The number of blank TR between stimulus and feedback was jittered using a truncated geometric distribution ($p = 0.5$; max TR = 3) (Ashby, 2019), and the number of blank TR between feedback and the next stimulus was also jittered using a truncated geometric distribution ($p = 0.5$; max TR = 5). When more than 1 blank TR was presented between feedback on trial t and the stimulus on trial $t+1$, a fixation cross was shown for 1 TR immediately before stimulus presentation (replacing the last blank TR, so that there was always at least 1 blank TR between the feedback and the next stimulus). The crosshair was shown on an average of 48% of the trials. A schematic showing a trial in each condition is shown in Figure 2.

The test phase (block 6) was identical to the training blocks except that participants were now asked to separate "B" stimuli from "C" stimuli. Specifically, the question "B or C?" was used in every test trial of the A/B condition. In the YES/NO condition, test trials used the questions "Is this a 'B'?" or "Is this a 'C'?", and correct "NO" responses were always from the "C" or "B" categories (respectively). Cate-

[1] The texture was one of the masks used in Hélie and Cousineau (2015). For an example texture, see Figure 1 in the cited article.

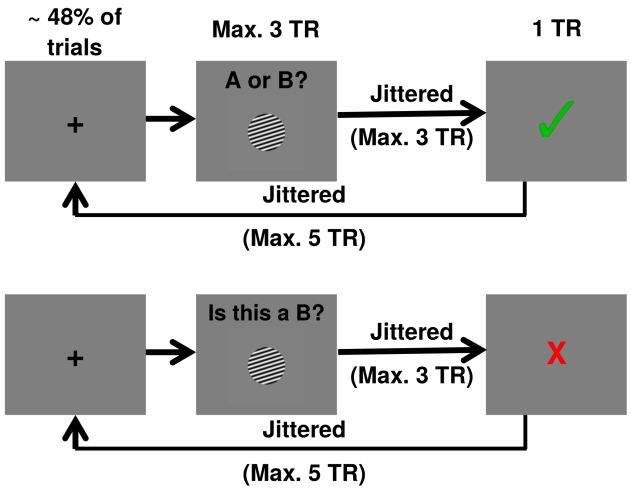

Figure 2. Experimental procedures. The top row shows an example A/B trial while the bottom row shows an example YES/NO trial. 1 TR = 720 ms.

gories and questions were equally likely. A non–informative texture replaced the feedback during the test phase.

Neuroimaging

A rapid event–related design fMRI procedure was used to examine BOLD signal as participants categorized visual stimuli. The scanning session was conducted at the Purdue Life Science MRI Facility using a 3T Siemens Prisma scanner with a 64–channel head coil. Each block in the experiment used a separate scan. Functional runs used the lifespan HCP multiband echo–planar images (EPI) sequence (S. M. Smith et al., 2013). The sequence parameters were as follows: multiband acceleration factor: 8; TR: 720 ms; echo time (TE): 30 ms; flip angle (FA): 52°; field of view (FOV): 210 mm. Each volume consisted of 72 slices acquired parallel to the static magnetic field (z–direction) and each slice was a matrix of 104×104. The resulting voxels where 2 mm isometric. Each functional run had a different number of TR (because of jittering) but lasted about 8 minutes. Before the beginning of the experiment, a regular localizer was run, and after the experiment a T1–weighted MPRAGE (TR = 2,300 ms; TE = 2.98 ms; FA = 9°; 176 sagittal slices; 1.1 mm thick; 1 mm × 1 mm in–plane resolution; 256×256 matrix) high–resolution structural scan was run. Each scanning session lasted about 60 minutes. The experimenter talked with the participant between each scan, and the participant was allowed to take a break between each scan (but not to exit from the scanner). These manipulations were designed to minimize fatigue and monotony.

Computational modeling

Computational models were fit to the categorization responses of individual participants in the last two blocks of training and used to predict each participant's categorization responses during the test block and identify the type of category representation that was learned. Hélie, Shamloo, and Ell (2017) showed that the type of representations learned by individual participants can be identified by fitting density–based and boundary–based models to their responses. Specifically, we fit Gaussian density models to capture within–category representations and linear boundary–based models to capture between–category representations. These models have been selected because the optimal bounds separating the categories used in the experiment are linear, and the optimal bound separating two Gaussian distributions is also linear – so both models are optimal for the categories used. Also, the density models are generative models that rely on each category's statistics (e.g., category mean and covariance), which are examples within–category information. In contrast, the boundary–based models are classification models that rely on separation bounds and noise on the separation bounds. These are examples of betweeen–category information.

Because of the information contained in the models (i.e., what the model parameters represent), when the best–fitting model is density–based the participant is said to have learned within–category information. In contrast, participants best–fit by boundary–based models are said to have learned between–category information. Modelling details and fitting procedures are described in the Appendix.

Neuroimaging analysis

Preprocessing and data analysis were conducted using FEAT (FMRI Expert Analysis Tool) version 6.00, part of FSL (fsl.fmrib.ox.ac.uk/fsl/fslwiki). Preprocessing was done separately on each EPI scan to reduce sources of noise and artifact, including motion correction using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), BET brain extraction (S. Smith, 2002), and spatial smoothing with a FWHM of 4 mm and a high pass temporal filter with a cutoff of 100 seconds. Each functional scan (EPI) was linearly aligned with the participant's structural scan and a non–linear transformation was used for normalization to the `MNI152_2mm_brain` template. Scanning data with excessive head motion (i.e., greater than 1 mm) was micro–scrubbed by creating a nuisance regressor for each TR in which motion was excessive.

Main task–based analysis. First, low–level analyses were performed separately on each EPI scanning block. Three events were defined: stimulus, feedback, and blank. The stimulus events from error trials were not included in all neuroimaging analyses. Each event was modeled by a sepa-

rate regressor and convolved with a double–gamma haemodynamic response function. A temporal derivative and temporal filtering were added to the design matrix. The contrast of interest was Stimulus > Blank, and Feedback was used as a nuisance regressor. Second, the results of the low–level analyses were input into mid–level analyses to aggregate the training block data (Blocks 1–5). The mid–level analyses yielded a separate brain map for each participant. Both the low– and mid–level analyses used fixed effects modeling. Next, a high–level analysis was performed using the output of the mid–level analysis as input (i.e., one aggregated brain map of training blocks for each participant). Two contrasts were calculated: (1) A/B > YES/NO and (2) YES/NO > A/B. These contrasts should reflect task characteristics (regardless of category representation). The high–level analysis used random effects modeling (FLAME 1+2) with a threshold value of $Z > 1.96$ and cluster–size correction of $p < .05$. While this Z threshold can be considered lenient, it has been used in earlier categorization work and Eklund, Nichols, and Knutsson (2016) have shown that using FLAME with cluster correction does not yield as many false positives as other reviewed methods when using a rapid event–related design.

Exploratory analyses. In addition to the main analysis described in the previous section, three exploratory analyses were performed: (E1) a model–based training analysis, (E2) a task–based training functional connectivity analysis, and (E3) a task–based test analysis.

The first exploratory analysis (E1) used a *post hoc* classification of the participants based on the fit of computational models to the behavioral data (Hélie, Shamloo, & Ell, 2017). Computational models were fit to the categorization responses of individual participants in the last two blocks of training and used to predict the participant's categorization responses during the test block and identify the type of category representation that was learned. However, the computational models used do not allow for traditional model–based analyses (Ashby, 2019). The problem is model mimicry. During training, both the density–based and boundary–based models can mimic each other. The test phase was designed specifically to allow for distinguishing between the models (for details see the Appendix). The index used for model selection was the accuracy of the model fit to the training data on each participant's test data (i.e., cross–validation error). Crucially, the test data were not used to estimate the model parameters. As a result, the model fit is scaled between 0 and 1 (1 meaning that the model perfectly predicts test data and 0 meaning that it cannot predict a single test trial).

For the model–based fMRI analyses we calculated a weighed average of the BOLD signal of all participants weighed by their model fit. For example, one participant may have a fit of 0.50 for the boundary model and a fit of .75 for the density model. To calculate the weights, we normalized these fits so that this participant's weight is 0.5 / (0.5 + 0.75) = 0.4 for the boundary model and 0.75 / (0.5 + 0.75) = 0.6 for the density model (for each participant, model weights sum to 1). For the density model–based fMRI analysis, we used the output of the mid–level analyses described above as input and calculated the mean for all non–random participants, using that participant's density fit as a weight. The same was done for the boundary model–based fMRI analysis. We then compared these means: (1) Boundary > Density and (2) Density > Boundary. We used the same random effects modeling, thresholds, and corrections as in the main task–based analysis described above. In addition, we also performed a conjunction analysis on the mean Density and Boundary BOLD signal to identify a shared categorization network. The conjunction analysis used a threshold value of $Z > 3.72$ and cluster–size correction of $p < .05$.

The second exploratory analysis (E2) was used to assess task–based functional connectivity during training. We used a PPI analysis with a seed in the MTL. First, a MTL mask was created by adding the following anatomical masks in MNI space: 1) Parahippocampal Gyrus, anterior division and 2) Parahippocampal Gyrus, posterior division (both from Harvard-Oxford Cortical Structural Atlas); 3) Left Hippocampus and 4) Right Hippocampus (both from Harvard-Oxford Subcortical Structural Atlas); 5) GM hippocampus entorhinal cortex L and 6) GM Hippocampus entorhinal cortex R (both from Juelich Histological Atlas). The resulting mask was then binarized and transformed to each individual participant's structural space. The following was repeated individually for each participant training block (Blocks 1–5). First, the mean time series within the mask was calculated. This was used as the physiological regressor. Second, the stimulus event was used as the psychological regressor. A third regressor was created as the interaction of the first two. The same preprocessing steps were used as for the regular BOLD analyses described above, and the low–level PPI analyses were input into mid–level analyses to calculate an average PPI brain map for each participant. Again, both the low– and mid–level analyses used fixed effects modeling. Finally, the mid–level analyses were input into a high–level analysis that used random effects modeling (FLAME 1+2) and calculated the same contrasts using the same thresholds as the main task–based analysis described above.

The third exploratory analysis (E3) was a task–based analysis on the test data. This analysis was identical to the main task–based analysis, except that the input was the output of the low–level analysis performed on the test block (Block 6). This last analysis was used to address the exploratory aim of whether the removal of feedback and new category comparison would affect the brain signal responsible for categorization decisions. This last exploratory analysis also used random effects modeling (FLAME 1+2) and calculated the same contrasts using the same thresholds as the main task–based analysis described above.

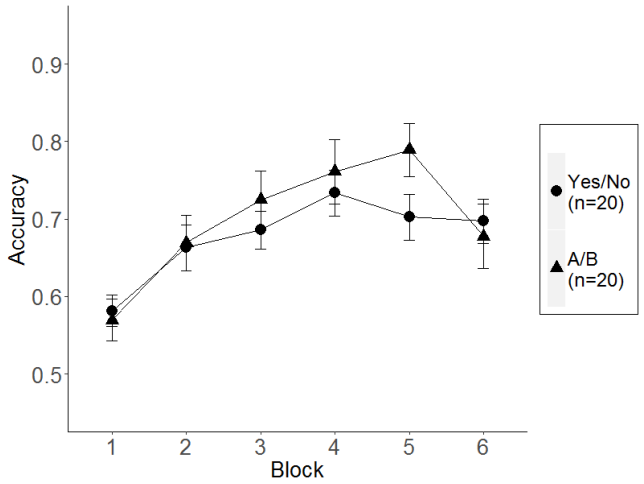

Figure 3. Mean accuracy per block. Blocks 1–5 are the training phase and Block 6 is the test phase. Error bars are between–subject standard error of the mean.

Results

The data from two participants in the A/B condition were removed for excessive head motion (> 1 mm) over the entire session (so that micro–scrubbing was not possible). The following analyses thus include the data from 20 participants in each conditon.

Behavioral results

The mean accuracy for each block in each condition is shown in Figure 3. As can be seen, training accuracy was similar for the A/B and YES/NO training tasks. This was confirmed by a Task (A/B, YES/NO) × Training Block (1...5) mixed ANOVA. The effect of Block was statistically significant ($F(4, 152) = 31.41, p < .0001, \eta^2 = 0.17$) but the effect of Task ($F(1, 38) = 0.59, p = .4474, \eta^2 = 0.01$) and the interaction between the factors ($F(4, 152) = 2.18, p = .0736, \eta^2 = 0.01$) were not. The mean accuracy in Block 1 was 57.5%, which improved to 74.5% in Block 5, and no difference was found between the learning rate in both tasks.

Transfer cost, however, differed between the two tasks. Categorical knowledge was better transferred in the YES/NO task than in the A/B task. This was confirmed by a Task (A/B, YES/NO) × Block (Last training block, Test block) mixed ANOVA. The effect of Block reached statistical significance ($F(1, 38) = 6.13, p = .0179, \eta^2 = 0.04$) while the effect of Task did not ($F(1, 38) = 0.64, p = .4274, \eta^2 = 0.01$). However, these effects need to be interpreted in the context of a statistically significant interaction ($F(1, 38) = 5.07, p = .0303, \eta^2 = 0.03$). Since our interest in is transfer cost, we decomposed the effect of Block within each level of Task, and the results show a statistically significant cost in performance at test for the A/B task ($t(19) = 3.68, p = .0016, \eta^2 = 0.42$) but not for the YES/NO task ($t(19) = 0.15, p = .8853, \eta^2 = 0.00$). The mean accuracy cost in the A/B task was 11.1% while the mean accuracy cost in the YES/NO task was 0.5%. Both the training and transfer ANOVA reproduced the results in Hélie, Shamloo, and Ell (2017) Experiment 1.

Model–based results. Boundary–based models, density–based models, and a random response model were fit to the participants' data. In the YES/NO task, 7 participants were best–fit by the density–based model, 8 participants were best–fit by the boundary–based model, and the remaining 5 participants were best–fit by the random model. In the A/B task, 10 participants were best–fit by the density–based model, 6 participants were best–fit by the boundary–based model, and the remaining 4 participants were best–fit by the random model. We did not find any evidence that the training task affected the distribution of best–fitting models ($\chi^2(2) = 0.93, p = .6293$). This result is different from Hélie, Shamloo, and Ell (2017) Experiment 1, who found more participants best–fit by the boundary model in the A/B training task condition.

Neuroimaging results

Main task–based analysis. The training results for task–related contrasts are listed in Table 1. As can be seen, five clusters have been identified for the contrast YES/NO > A/B. These clusters are shown in Figure 4 (red/yellow). As discussed in the *Introduction* section, one important difference between the YES/NO task and the A/B task is that with A/B there are direct Stimulus → Response associations. For example, if a given stimulus is a member of category "A", then the left button in the left hand is always the correct button press. Likewise, a "B" stimulus is always associated with the right button in the left hand. However, this is not the case with the YES/NO task. A member of category "A" is sometimes associated with a button press in the left hand ("YES") and sometimes associated with a button press in the right hand ("NO"), depending on which question was asked. As a result, rules learned in the A/B task are univalent (i.e., each stimulus is associated with a unique button press) whereas rules learned in the YES/NO task are bivalent (i.e., each stimulus can be associated to more than one button press, depending on context) (Bunge, 2004). Clusters 2, 4, and 5 are consistent with this task difference. These clusters include the preSMA, ventrolateral PFC, and frontal pole. The preSMA is involved in internally generated movement (Nachev et al., 2008), which happens when participants need to choose which button to press after having categorized the stimulus. The ventrolateral PFC is an important part of the WM network (Ashby et al., 2005) and is involved in rule application (Tsujii, Masuda, Akiyama, & Watanabe, 2010). The frontal pole plays an important role in higher–level cognition and set–shifting (Simard et al., 2011;

Wang et al., 2010). Clusters 1 and 3 include the angular and supramarginal gyri. The right angular gyrus has been associated with reward acquisition while interacting the environment (S. W. Cole, Yoo, & Knutson, 2012). The left supramarginal gyrus has been associated with visual feature integration and coordination in WM (Morgan et al., 2011). So these two clusters may be supporting different, complementary function. On the one hand, the right angular gyrus is also strongly connected with the posterior hippocampus, which is thought to contain detailed, fine–grained stimulus representations (Bowman & Zeithamova, 2018). Because the YES/NO task requires answering questions about specific stimuli, a more detailed representation may be critical and allow for better estimating the likelihood of obtaining a reward in each trial (e.g., by estimating the distance of the current stimulus from a prototype). On the other hand, the left supramarginal gyrus has been observed in task–switching experiments (Philipp et al., 2013). This result, along with the findings for C2, is consistent with the hypothesis that participants may use the question as context to select the appropriate rule and treat each question as a separate task. Such operation would require coordination in WM (Fleischer & Hélie, 2020). While speculative, if this interpretation is correct participants would switch task when the question changes.

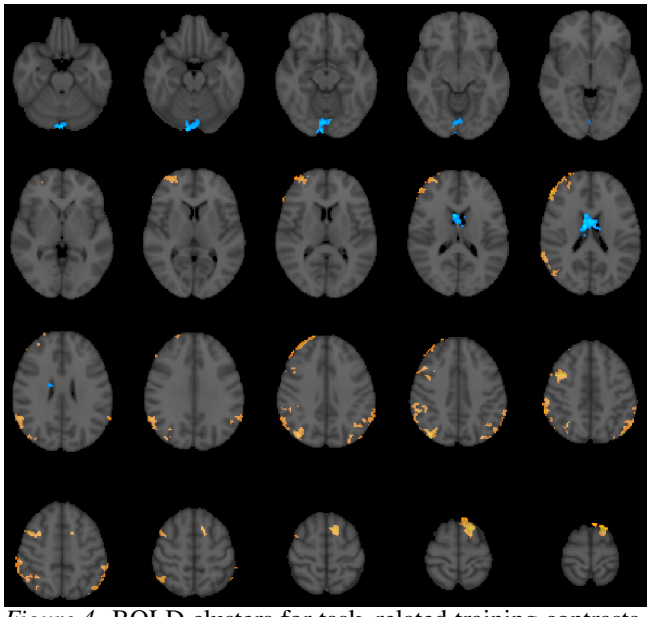

Figure 4. BOLD clusters for task–related training contrasts. The slices shown range from $z = -25$ to $z = 75$ from left to right and top to bottom by jumps of 5. YES/NO > A/B clusters are shown in red/yellow while A/B > YES/NO are shown in blue. Cluster coordinates are listed in Table 1.

Table 1 and Figure 4 also show two clusters that were more activated in the A/B task than in the YES/NO task (blue). Cluster 7 spanned the thalamus and caudate nucleus, which are part of a network of areas related to associative learning (Hélie, Ell, & Ashby, 2015). For example, the COVIS model of category learning (Ashby et al., 1998) assigns these brain areas to procedural learning, and Hélie et al. (2010) found similar brain activity for initial learning of Stimulus → Response associative rules. Hélie et al. suggested that early learning of Stimulus → Response associations may be critical for succesful hypothesis–testing and rule learning (see also Hélie, Proulx, & Lefebvre, 2011).

Overall, the task–based results support the hypothesis that participants in the YES/NO task may use bivalent rules and switch task as a function of the question displayed on the screen. In contrast, the BOLD signal in the A/B condition is more consistent with participants learning direct Stimulus → Response rules.

Exploratory analysis 1: Model–based training. The first exploratory analysis was an attempt at identifying BOLD signal related to the participants' type of category representation. Computational models were fit to the participant's behavioral data and the model fits were used to weigh each participant into average brain maps. As a reminder, participants best–fit by the density model were thought to have learned within–category information, while participants best–fit by the boundary model were thought to have learned between–category information. Participants best–fit by a random model were not included in this exploratory analysis. Three contrasts were calculated: (1) Boundary & Density, (2) Boundary > Density, and (3) Density > Boundary. Contrast (1) was a conjunction map that should reveal a joint categorization network, whereas contrasts (2) and (3) should show representational differences. The resulting clusters are listed in Table 2 and shown in Figure 5. Note that no cluster survived correction for multiple testing for Density > Boundary.

As can be seen, the conjunction analysis yielded 14 clusters (shown in Blue in Figure 5). Cluster 1 was very large and the coordinates of local maxima are shown in Table 3. Most of the maxima are located in the temporal occipital fusiform and lateral occipatal cortices. Other clusters listed in Table 2 include the middle frontal gyrus (C2, C3, C13), orbitofrontal / insular cortex (C4), the thalamus (C6, C7, C9), and the caudate nucleus (C8, C11). This corresponds to a typical category learning network (Carpenter, Wills, Benattayallah, & Milton, 2016; Hélie et al., 2010; Milton, Bealing, Carpenter, Bennattayallah, & Wills, 2017; Seger, Dennison, Lopez-Paniagua, Peterson, & Roark, 2011; Seger, Braunlich, Wehe, & Liu, 2015; Zeithamova et al., 2008, 2019). It is interesting to note that this network is observed even when results from two seperate tasks are pooled together. This suggests that these brain areas may be related to category learning generally and not to the specific tasks used in the laboratory.

Figure 5 also shows clusters for the Boundary > Density contrast (red/yellow). Two clusters were statistically significant (see Table 2). The first cluster (C15) span the right angu-

Table 1
Task–related BOLD clusters during training

Cluster	Size	*p*	*Max*(*Z*)	Coordinates			Brain regions
				x	*y*	*z*	
YES/NO > A/B							
1	1,126	< .0001	4.22	48	-56	54	R. angular gyrus
2	639	< .0001	4.09	34	58	10	R. frontal pole
3	549	.002	3.88	-58	-40	54	L supramarginal gyrus
4	429	.009	4.45	-20	14	68	L. superior frontal gyrus
5	348	.036	3.95	38	6	44	R. middle frontal gyrus
A/B > YES/NO							
6	354	.033	3.52	10	-90	-18	L./R. lingual gyrus
7	346	.037	3.86	6	8	18	R. thalamus/caudate

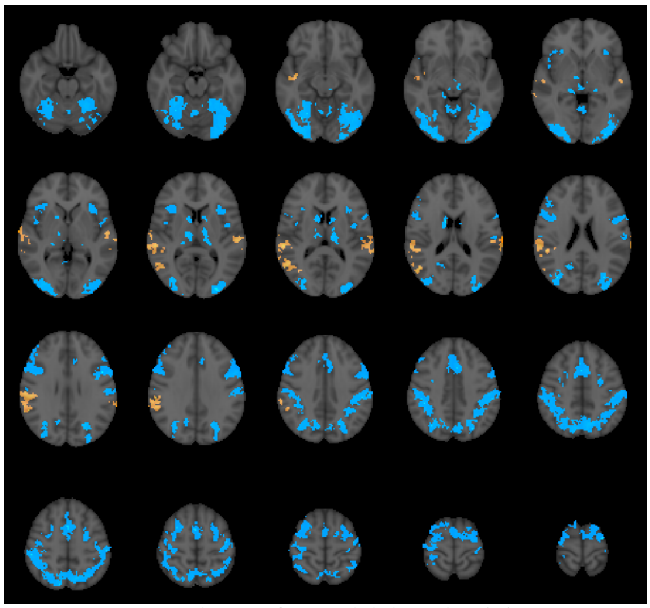

Figure 5. BOLD clusters for model–based training contrasts. The slices are the same as in Figure 4. Boundary > Density clusters are shown in red / yellow, while clusters from the conjunction analysis are shown in shades of blue. No cluster survived correction for the Density > Boundary contrast. All cluster coordinates are listed in Table 2.

lar gyrus and the superior temporal gyrus. These brain areas are part of posterior parietal cortex (as hypothesized). While right angular gyrus activation was also found in the YES/NO > A/B contrast (see Table 4), C15 is more anterior and inferior and there is little overlapping activation between these clusters. Previous research linked activity in C15 to cognitive control, specifically response inhibition in the stop–signal–task (Boehler, Appelbaum, Krebs, Hopf, & Woldorff, 2010; Hwang, Velanova, & Luna, 2010). It is unclear why participants learning between–category information would show more activity related to response inhibition. One intriguing possibility is Hélie et al. (2015) proposed that perceptual categorization rules were implemented by inhibiting incorrect motor programs using pre–synaptic inhibition (instead of activating the correct motor program). In Hélie et al.'s computational model, however, the pre–synaptic inhibition was implemented in lateral PFC – not posterior parietal cortex.

Exploratory analysis 2: Task–based training functional connectivity. The second exploratory analysis used PPI on the training data with a MTL seed. The resulting clusters of activity are listed in Table 4 and shown in Figure 6. As can be seen, 5 clusters reached statistical significance in the YES/NO > A/B contrast (red/yellow) and 2 clusters reached statistical singificance in the A/B > YES/NO contrast (blue). For the YES/NO task, MTL functional connectivity was increased in a wide network that span the right superior parietal lobule (C1) and middle/superior frontal gyri (C2, C5). These clusters overlap with the dorsal attention network and a previously identified rule learning network (Bowman & Zeithamova, 2018; Zeithamova et al., 2019). These clusters also overlap with the cognitive control network (M. W. Cole & Schneider, 2007). This is consistent with the hypothesis that participants used bivalent rules in the YES/NO task. Considering context in response selection may put more demands on WM and require a higher level of cognitive control (Majerus et al., 2017) than the univalent rules used in the A/B task. This interpretation makes intuitive sense given (1) that participants see a larger number of possible questions in this task (4 in the YES/NO task vs. 2 in the A/B task) and (2) the absence of a direct Stimulus → Response association. Cluster C3 is also noteworthy as it includes the left angular/supramarginal gyrus. Note that activity in this brain area was also observed in the YES/NO

Table 2
BOLD clusters during training for contrasts based on category representation

				Coordinates			
Cluster	Size	p	$Max(Z)$	x	y	z	Brain region(s)
Conjunction							
1	16,265	< .0001	6.14	36	-50	-16	R. temporal occipital fusiform complex
2	816	< .0001	5.42	50	14	26	R. middle frontal gyrus
3	801	< .0001	5.98	-48	8	28	L. middle frontal gyrus
4	307	< .0001	5.39	36	26	0	R. frontal orbital / insular cortex
5	205	< .0001	5.55	-28	18	10	L. insular cortex / frontal operculum
6	196	< .0001	4.56	-10	-20	10	L. thalamus
7	176	< .0001	4.6	-8	-20	-6	L. thalamus / brainstem
8	160	< .0001	5.55	12	2	18	R. caudate
9	122	< .0001	4.57	14	-16	6	R. thalamus
10	98	< .0001	4.3	-34	-4	16	L. insular cortex / central operculum
11	62	.0004	4.4	-10	4	14	L. caudate
12	59	.0006	4.53	-2	-56	-36	L. cerebellum
13	36	.0129	4.07	44	26	38	R. middle frontal gyrus
14	30	.0323	4.4	-22	-34	-44	L. cerebellum
Boundary > Density							
15	1328	< .0001	4.23	60	-46	26	R. angular gyrus / supramarginal gyrus / parietal operculum
16	390	.0249	3.23	-56	-22	6	L. parietal operculum / planum temporale
Density > Boundary							
None.							

> A/B contrast (see Table 4), but there is no overlap. Activity in the cluster found in the PPI analysis has been associated to memory encoding (Vaidya, Zhao, Desmond, & Gabrieli, 2002) and retrieval (Stock, Röder, Burke, Bien, & Rösler, 2009), and may play an important role in category representation.

Table 4 and Figure 6 also show cluster activity for the A/B > YES/NO contrast. As can be seen, two clusters reached statistical significance, both located in the PFC. The frontal pole is related to decision–making, reward processing (Hélie, Shamloo, Novak, & Foti, 2017; O'Doherty, Cockburn, & Pauli, 2017), learning bivalent rules and rule sets (Badre, 2008; Fleischer & Hélie, 2020), and memory generalization (Zeithamova et al., 2019). Relatedly, the dorsolateral PFC is related to cognitive control and WM (and is part of both the dorsal attention network and the cognitive control network; Ashby et al., 2005; M. W. Cole & Schneider, 2007; Vossel et al., 2014). This result suggests that different parts of the dorsal attention network may be involved in the two learning tasks, possibly corresponding to encoding load and cognitive control (Majerus et al., 2017). Interestingly, these brain areas showed more task–related activity in the YES/NO task. Yet, results from the exploratory PPI analysis suggest that their task–based functional connectivity with the MTL is stronger for the A/B task. As a reminder, the PPI analysis was exploratory, and more research is needed to better understand the possibly different role of the dorsal attention in these two training tasks.

Exploratory analysis 3: Task–based test. The last exploratory analysis tested contrasts focused on Block 6. While

Table 3
Task–related local maxima in Cluster 1 from Table 2

		Coordinates			
Maximum	*Max*(*Z*)	*x*	*y*	*z*	Brain regions
1	6.14	36	-50	-16	R. temporal occipital fusiform cortex
2	6.07	30	-46	-20	R. temporal occipital fusiform Cortex
3	6.06	50	-60	-6	R. inferior temporal gyrus
4	5.72	-24	-86	4	L. lateral occipital cortex
5	5.68	-28	-94	8	L. occipital pole
6	5.67	-20	-66	34	L. lateral occipital cortex
7	5.66	30	-84	14	R. lateral occipital cortex
8	5.66	-30	-92	-4	L. occipital pole
9	5.65	28	-60	-14	R. temporal occipital fusiform cortex
10	5.63	26	-56	-14	R. temporal occipital fusiform cortex
11	5.61	36	-76	22	R. lateral occipital cortex
12	5.61	38	-74	36	R. lateral occipital cortex
13	5.60	34	-68	-12	R. occipital fusiform gyrus
14	5.60	-30	-48	-16	L. temporal occipital fusiform cortex
15	5.60	-22	-98	-10	L. occipital pole
16	5.58	32	-46	-34	R. cerebellum
17	5.57	-32	-80	24	L. lateral occipital cortex
18	5.56	-20	-66	38	L. lateral occipital cortex
19	5.56	32	-50	-34	R. cerebellum
20	5.55	30	-48	-24	R. cerebellum

Table 4
Task–related PPI clusters during training with MTL seed

				Coordinates			
Cluster	Size	*p*	*Max*(*Z*)	*x*	*y*	*z*	Brain region(s)
YES/NO > A/B							
1	2035	< .0001	3.86	28	-68	48	R. superior parietal lobule
2	1127	< .0001	4.54	52	16	46	R. middle frontal gyrus
3	490	.0011	3.49	-44	-54	50	L angular/supra-marginal gyrus
4	323	.0236	3.51	52	-76	-2	R. lateral occipital cortex
5	310	.0305	3.47	-26	18	56	L. superior frontal gyrus
A/B > YES/NO							
6	1704	< .0001	4.48	-34	42	14	L. frontal pole
7	323	.0236	4.26	-16	24	40	L. superior frontal gyrus

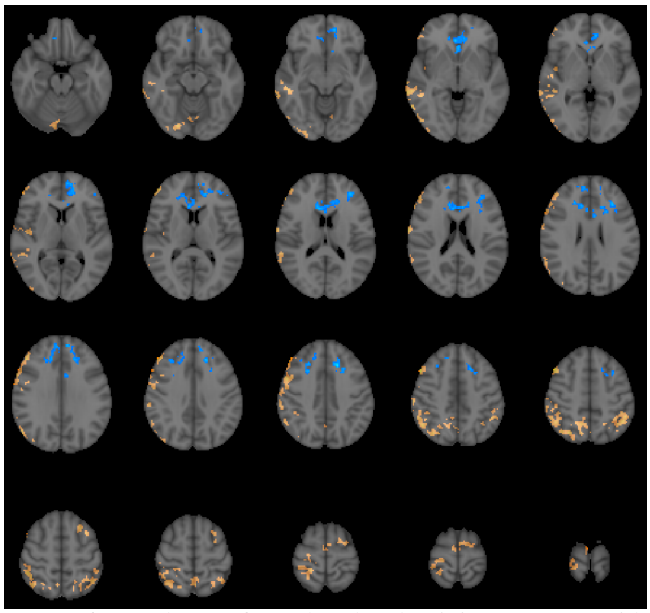

Figure 6. PPI clusters for task–related training contrasts with MTL seed. YES/NO > A/B clusters are shown in red/yellow while A/B > YES/NO are shown in blue. Cluster coordinates are listed in Table 4.

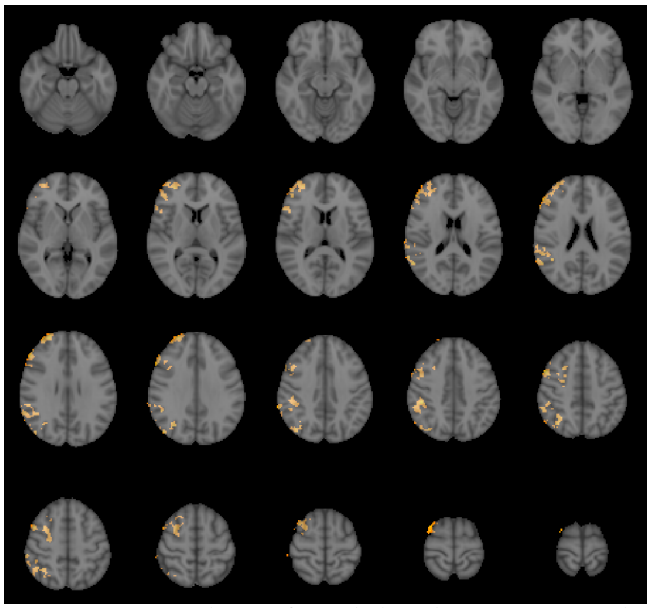

Figure 7. BOLD clusters for task–based test contrasts. The slices are the same as in Figure 4. YES/NO > A/B clusters are shown in red/yellow. There was no statistically significant cluster for A/B > YES/NO. Cluster coordinates are listed in Table 5.

the test block was initially included to eliminate model mimicry and allow for classifying participants based the type of knowledge representation learned, the data might contain interesting information about the possible effects of removing feedback. As a reminder participants performed the same task at test as they did at training using the same categories. However, new pairs of categories were contrasted (i.e., "B" vs. "C"). The other difference is that accuracy feedback was no longer provided.

Statistically significant clusters found at test are listed in Table 5. As can be seen, three clusters were found for the YES/NO > A/B contrast, and none were found in the A/B > YES/NO contrast. The YES/NO > A/B clusters are shown in Figure 7. One cluster was located in the right middle PFC while another was located in the right intraparietal sulcus. These clusters are very close to C1 and C5 observed in training (see Table 1), which suggests that participants in the YES/NO task may be using a similar response strategy to what they used during training, namely using bivalent rules and switching task using the question as a context cue.

Discussion

The present experiment explored the effects of training methodology and category representation on brain activity in a rule–based categorization task. The main findings are: (1) participants in the YES/NO training task may learn contextual rules and treat each contextual question as a different task, and (2) individual differences in the type of category information learned can produce different BOLD response. These novel findings have important implication for future research in categorization that will now be explored in turns.

What type of rules are learned in categorization?

Previous empirical (for a review, see Ashby & Valentin, 2017) and computational modeling (Ashby et al., 1998) work suggest that participants can learn verbalizable classification rules using hypothesis–testing. Rules can be univalent (each stimulus is always associated with the same motor response) or bivalent (each stimulus can be associated with more than one motor response depending on context) (Bunge, 2004). Rules can contain between–category information (e.g., mice are smaller than cats) or within–category information (e.g., mice are small) (Hélie, Shamloo, & Ell, 2017). Most previous work on category learning has treated the YES/NO categorization task as equivalent to the A/B categorization task. One exception was provided in Zeithamova et al. (2008), who used high–dimensional stimuli with discrete dimensions and showed that performance in the A/B task was mediated by brain areas associated with episodic memory whereas performance in the YES/NO task was mediated by brain areas associated with non–declarative memory. These results were surprising given that the A/B task requires learning univalent rules whereas the YES/NO task requires learning bivalent rules. Bivalent rules are typically represented more rostrally in the PFC (Badre et al., 2010; Bunge, 2004; Hélie et

Table 5
Task–related BOLD clusters during test

				Coordinates			
Cluster	Size	*p*	*Max*(*Z*)	*x*	*y*	*z*	Brain region(s)
YES/NO > A/B							
1	1866	< .0001	4.06	36	10	66	R. middle frontal gyrus
2	778	.0001	3.56	46	-34	42	R. intraparietal sulcus
3	384	.0278	3.52	46	-78	32	R. lateral occipital cortex
A/B > YES/NO							
None.							

al., 2010). One possibility is that, since the stimuli in Zeithamova et al. had binary dimensions and were more distinct from each other, participants were able to memorize the Stimulus → Response associations and retrieve them from memory.

In the present experiment, each stimulus was unique and the stimulus dimensions varied continuously, so memorizing the stimuli was not a viable strategy. For example, Hélie et al. (2010) used similar sine–wave gratings with the A/B task and showed that hippocampus activity was negatively correlated with categorization accuracy. To perform well in this task, participants needed to learn rules that can be generalized to new stimuli. As a result, the bivalent rules that participants needed to learn in the YES/NO task produced more activity in the preSMA, ventrolateral PFC, and frontal pole. They also showed increased task–based functional connectivity between the MTL and the dorsal attention network, as well as between the MTL and the cognitive control network. These results are consistent with the hypothesis that contextual rules require more attentional resources and cognitive control. Interestingly, brain areas related to task–switching also showed increased activity in the YES/NO task, consistent with the possibility that participants may have treated each question as a separate task and switched task when the question changed. This interpretation is speculative, and more research is needed to confirm its correctness.

In contrast, univalent rules were sufficient in the A/B task, which produced more activation in the thalamus and caudate. These areas form a network that is well–suited for Stimulus → Response associative learning (Hélie et al., 2015). Importantly, the question asked did not affect which button needed to be pressed after category membership was decided. There was thus no need to treat each question as a separate task (Fleischer & Hélie, 2020). Overall, the qualitative differences in the type of rules that participants needed to learn in the A/B and YES/NO categorization tasks were well reflected by the task–based fMRI results.

What type of information is learned in categorization?

Hélie, Shamloo, and Ell (2017) showed that some participants learn rules containing within–category information while others learn rules containing between–category information. The type of category representation that participants learn can be identified using computational modeling. In the present experiment, we found that about half the non–random participants in each task condition learned each type of category representation. Unsurprisingly, the results show that there is a large network of category learning brain areas that is shared by participants learning both types of representations (as reviewed in Zeithamova et al., 2019), but there were also some differences. Specifically, two clusters of activity were found in the anterior lateral portion of inferior parietal cortex for participants learning between–category information. Stillesjö, Nyberg, and Wirebring (2019) argued that inferior parietal cortex is associated with similarity–based processes, and Zeithamova et al. (2019) reviewed evidence showing that this brain area is related to category representation in both human and non–human primates. Specifically, the inferior parietal cortex has been related to the probability of receiving a reward. In the context of between–category information, this could correspond to the distance–to–bound effect, which shows that the probability of receiving a reward increases as the stimulus is located further away from the decision bound. Hélie, Waldschmidt, and Ashby (2010) have shown that the distance–to–bound effect is reduced with extensive practice, so if this interpretation is correct these clusters of activity would also diminsh with extensive practice. The present experiment was not designed to directly test for this possibility, so more research is needed to verify this interpretation.

Limitations and future work

One important limitation of the present experiment is sample size. The experiment included 40 participants, which is typically sufficient in most contexts but can be small when

participants are separated into subgroups using computational modeling. As a result, the present study was not able to reproduce the bias towards between–category representation previously observed for rule–based categories in the A/B task (Hélie, Shamloo, & Ell, 2017). The pattern of responses associated with each type of category representation was well reproduced, but some of the brain activity related to each type of category representations may have been missed. Future research should include more participants, which would also allow for testing possible Task (A/B, YES/NO) × Category Representation (Boundary, Density) interactions.

Another important limitation is related to the duration of the test phase. The computational models used to identify the type of category representations learned by the participants rely on categorization accuracy at test (Hélie, Shamloo, & Ell, 2017). It is possible that the observed performance at test is transitory, and that categorization would improve and accuracy cost diminish if the test was longer. A possible change in performance could affect the identification of the type of representations learned by the participants, which could affect the BOLD signal. Future research should include a longer test period, perhaps as long as the training phase, to confirm that 100 test trials is sufficient to accurately identify the type of category representations learned by individual participants.

Open Practices Statement

The data and materials for the experiment are available by contacting the corresponding author. The experiment was not preregistered.

Acknowledgements

Correspondence should be addressed to Dr. Sébastien Hélie, shelie@purdue.edu. The authors declare no competing financial interests. This research was supported in part by NSF grant #1349677-BCS and NIMH grant #2R01MH063760 to SH, as well as NSF grant #1349737 to SWE. The authors would like to thank Madison Fansher for help with data collection and analysis, as well as Aameneh Kermani and Stella Yao for help with data analysis.

References

Ashby, F. G. (2019). *Statistical Analysis of fMRI Data* (2nd ed.). Cambridge, MA: MIT Press.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., Ell, S. W., Valentin, V. V., & Casale, M. B. (2005). FROST: A distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, *17*(11), 1728–1743.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd Edition ed., pp. 157–188). Oxford: Elsevier.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal axis of the frontal lobes. *Trends in Cognitive Science*, *12*, 193–200.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, *66*, 315–326.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer.

Boehler, C. N., Appelbaum, L. G., Krebs, R. M., Hopf, J. M., & Woldorff, M. G. (2010). Pinning down response inhibition in the brain–conjunction analyses of the Stop–signal task. *NeuroImage*, *52*, 1621–1632.

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Bunge, S. A. (2004). How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 564–579.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathemtical Psychology*, *44*(1), 171–189.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A Comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*, 3557–3574.

Cole, M. W., & Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *NeuroImage*, *37*, 343–360.

Cole, S. W., Yoo, D. J., & Knutson, B. (2012). Interactivity and reward–related neural activation during a serious videogame. *PLoS One*, *7*, e33909.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*, 7900–7905.

Ell, S. W., Smith, D. B., Deng, R., & Hélie, S. (2020). Learning and generalization of within-category representations in a rule-based category structure. *Attention, Perception, & Psychophysics*, *82*, 2448–2462.

Ell, S. W., Smith, D. B., Peralta, G., & Hélie, S. (2017). The impact of category structure and training methodology on learning and generalizing within–category representations. *Attention, Perception, & Psychophysics*, *79*, 1777–1794.

Fleischer, P., & Hélie, S. (2020). A unified model of rule–set learning and selection. *Neural Networks*, *124*, 343–356.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*.

New York: Springer.

Hélie, S. (2006). An introduction to model selection. *Tutorials in Quantitative Methods for Psychology*, *2*, 1-10.

Hélie, S., & Ashby, F. G. (2012). Learning and transfer of category knowledge in an indirect categorization task. *Psychological Research*, *76*, 292-303.

Hélie, S., & Cousineau, D. (2015). Differential effect of visual masking in perceptual categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 816–825.

Hélie, S., Ell, S. W., & Ashby, F. G. (2015). Learning robust cortico-cortical associations with the basal ganglia: An integrative review. *Cortex*, *64*, 123–135.

Hélie, S., Ell, S. W., Filoteo, J. V., & Maddox, W. T. (2015). Criterion learning in rule–based categorization: Simulation of neural mechanism and new data. *Brain and Cognition*, *95*, 19–34.

Hélie, S., Proulx, R., & Lefebvre, B. (2011, apr). Bottom-up learning of explicit knowledge using a Bayesian algorithm and a new Hebbian learning rule. *Neural Networks*, *24*(3), 219–232. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/21239141` doi: 10.1016/j.neunet.2010.12.002

Hélie, S., Roeder, J., & Ashby, F. (2010). Evidence for cortical automaticity in rule-based categorization. *Journal of Neuroscience*, *30*(42), 14225–14234.

Hélie, S., Shamloo, F., & Ell, S. W. (2017). The effect of training methodology on knowledge representation in categorization. *PLOS ONE*, *12*, e0183904.

Hélie, S., Shamloo, F., Novak, K., & Foti, D. (2017). The roles of valuation and reward processing in cognitive function and psychiatric disorders. *Annals of the New York Academy of Sciences*, *1395*, 33–48.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, *72*(4), 1013–1031.

Hwang, K., Velanova, K., & Luna, B. (2010). Strengthening of top–down frontal cognitive control networks underlying the development of inhibitory control: a functional magnetic resonance imaging effective connectivity study. *Journal of Neuroscience*, *30*, 15535–15545.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*, 825–841.

Lee, A. C. H., Yeung, L. K., & Barense, M. D. (2012). The hippocampus and visual perception. *Frontiers in Human Neuroscience*, *6*, 91.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49–70.

Majerus, S., Péters, F., Bouffier, M., Cowan, N., & Phillips, C. (2017). The Dorsal Attention Network Reflects Both Encoding Load and Topdown Control during Working Memory. *Journal of Cognitive Neuroscience*, *30*, 144–159.

Markman, A. B., & Ross, B. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 529–613.

Milton, F., Bealing, P., Carpenter, K. L., Bennattayallah, A., & Wills, A. J. (2017). The neural correlates of similarity– and rule–based generalization. *Journal of Cognitive Neuroscience*, *29*, 150–166.

Morgan, H. M., Muthukumaraswamy, S. D., Hibbs, C. S., Shapiro, K. L., Bracewell, R. M., Singh, K. D., & Linden, D. E. (2011). Feature integration in visual working memory: parietal gamma activity is related to cognitive coordination. *Journal of Neurophysiology*, *106*, 3185–3194.

Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nature reviews. Neuroscience*, *9*(11), 856–869.

Nomura, E. M., & Reber, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience and Biobehavioral Reviews*, *32*(2), 279–291.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*, *68*(1), 73–100.

Philipp, A. M., Weidner, R., Koch, I., & Fink, G. R. (2013). Differential roles of inferior frontal and inferior parietal cortex in task switching: Evidence from stimulus–categorization switching and response–modality switching. *Human Brain Mapping*, *34*(8), 1910–1920.

Schneider, D. W., & Logan, G. D. (2014). Tasks, task sets, and the mapping between them. In J. A. Grange & G. Houghton (Eds.), *Task Switching and Cognitive Control* (pp. 27–44). New York: Oxford University Press.

Seger, C. A., Braunlich, K., Wehe, S., & Liu, Z. (2015). Generalization in category learning: The roles of representational and decisional uncertainty. *Journal of Neuroscience*, *35*, 8802–8812.

Seger, C. A., Dennison, C. S., Lopez-Paniagua, D., Peterson, E. J., & Roark, A. A. (2011). Dissociating hippocampal and basal ganglia contributions to category learning using stimulus novelty and subjective judgments. *NeuroImage*, *55*, 1739–1753.

Simard, F., Joanette, Y., Petrides, M., Jubault, T., Madjar, C., & Monchi, O. (2011). Fronto–striatal contribution to lexical set–shifting. *Cerebral Cortex*, *21*, 1084–1093.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype–based and exemplar–based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 800-811.

Smith, S. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*, 143–155.

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., ... for the WU-Minn HCP Consortium (2013). Resting–state fMRI in the Human Connectome Project. *NeuroImage*, *80*, 144–168.

Stillesjö, S., Nyberg, L., & Wirebring, L. K. (2019). Building memory representations for exemplar-based judgment: A role for ventral precuneus. *Frontiers in Human Neuroscience*, *13*(July), 1–16.

Stock, O., Röder, B., Burke, M., Bien, S., & Rösler, F. (2009). Cortical activation patterns during long-term memory retrieval of visually or haptically encoded objects and locations. *Journal of Cognitive Neuroscience*, *21*, 58–82.

Tsujii, T., Masuda, S., Akiyama, T., & Watanabe, S. (2010). The role of inferior frontal cortex in belief-bias reasoning: An rTMS study. *Neuropsychologia*, *48*(7), 2005–2008.

Vaidya, C. J., Zhao, M., Desmond, J. E., & Gabrieli, J. D. E. (2002).

Evidence for cortical encoding specificity in episodic memory: memory–induced re-activation of picture processing areas. *Neuropsychologia*, *40*, 2136–2143.

Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and ventral attention systems: Distinct neural circuits but collaborative roles. *Neuroscientist*, *20*(2), 150–159.

Wang, L., Liu, X., Guise, K. G., Knight, R. T., Ghajar, J., & Fan, J. (2010). Effective connectivity of the fronto–parietal network during attentional control. *Journal of Cognitive Neuroscience*, *22*, 543–553.

Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., van Kesteren, M. T., & Wutz, A. (2019). Brain mechanisms of concept learning. *Journal of Neuroscience*, *39*(42), 8259–8266.

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: Evidence from brain imaging and behavior. *Journal of Neuroscience*, *28*(49), 13194–13201.

Appendix

In this Appendix we describe the modelling procedure used to identify each participant's category representation. This procedure is adapted from Hélie, Shamloo, and Ell (2017). The first step was to identify participants who responded randomly at the end of training (i.e., non–learners). We used a Binomial distribution with $n = 100$ (number of trials in Block 5) and $p = 0.5$ (chance performance in each trial) to model a random responder in the last block of training. This distribution shows that the probability of obtaining an accuracy above 59% in Block 5 by responding randomly was less than $p < .05$ (bidirectional). As a result, participants with an accuracy below 59% in Block 5 were labeled as using the *random model*.

For all other participants, we fit Gaussian density models (e.g., Bishop, 2006) to capture within–category representations and linear boundary models (e.g., Maddox & Ashby, 1993) to capture between–category representations. These models have been selected because the optimal bounds separating the categories used in the experiments are linear, and the optimal bound separating two Gaussian distributions is also linear – so both models are optimal for the categories used.

Density model

The first step is to draw the participant's decision space. This means assigning a participant response to each coordinate point in the stimulus space (so that each participant has their own decision space). The result is similar to Figure 1 except that the participant's responses are used as symbols instead of the desired category labels. For the density model, a different gaussian distribution is fit to each possible response category (i.e., A–D). For YES/NO training, only trials in which the participant responded "yes" were included, because there is no way to know what category the participant had in mind when responding "no". For example, the "A" density is estimated by including only trials in which the question was "Is this an 'A'?" and the participant responded "yes". The same procedure was used to estimate the densities representing categories "B", "C", and "D". For A/B training, all trials in which the participant pressed the "A" response button were used to estimate the "A" density (or "B", or "C", or "D" to estimate the densities corresponding to categories B–D). In all cases, the maximum likelihood estimators were used (i.e., the sample mean and variance). Note that this model has 8 free parameters, i.e., the mean and variance of the radians for each category.

With the density model, the probability of identifying a stimulus as being located in the perceptual region associated with C_X is:

$$p(C_X|d_i) = \frac{f_X(d_i)}{f_X(d_i) + f_Y(d_i)} \tag{A1}$$

where $p(C_X|d_i)$ is the probability of locating stimulus d_i in the perceptual region associated with category X (denoted C_X), $f_X(d_i)$ is the probability of d_i according to the density estimated for C_X, and $f_Y(d_i)$ is the probability of d_i according to the density estimated for C_Y, where Y is the contrasting category. With the density model, Eq. A1 is used both at training and at test.

Boundary model

For the boundary models, the procedure is similar to that of the density models except that boundaries are estimated *between* the categories (instead of estimating densities *within* the categories). The procedure is exactly the same as described in (Maddox & Ashby, 1993). Each bound is represented by a gaussian distribution where the mean is the location of the bound and the variance corresponds to perceptual noise. The same participant's space as for the density model is drawn. Because the training phase only contrasted category "A" with "B" and category "C" with "D", only these two bounds were estimated. Estimating the AB boundary used all the trials in which the question referred to these categories for the YES/NO condition, and all the trials in which response buttons "A" or "B" were pressed for the A/B condition. The same procedure was used to estimate the CD boundary. All the parameters were estimated using maximum likelihood (Maddox & Ashby, 1993). Note that this model has 4 free parameters, corresponding to the location and noise of each bound.

With the boundary model, the probability of identifying a stimulus as being located on the A side of the AB bound during training is:

$$p(C_A|d_i) = 1 - F_{AB}(d_i) \tag{A2}$$

where $p(C_A|d_i)$ is the probability of locating stimulus d_i on the A side of the AB bound (denoted C_A), and $F_{AB}(d_i)$ is the probability of d_i according to the cumulative density function estimated for the AB bound. The probability of identifying a stimulus as being located on the B side of the AB bound is simply $p(C_B|d_i) = 1-p(C_A|d_i) = F_{AB}(d_i)$. The same equation applies at training for the CD bound (but substitute $C_C \rightarrow C_A$ and $C_D \rightarrow C_B$).

A different equation needs to be used at test with the boundary models because there is no BC bound, and both the AB and the CD bounds need to be considered in identifying a stimulus as being located in the C_B or C_C region. The boundary model at test is:

$$p(C_B|d_i) = \frac{F_{AB}(d_i)}{[1 - F_{CD}(d_i)] + F_{AB}(d_i)} \tag{A3}$$

where $p(C_B|d_i)$ is the probability of locating stimulus d_i in the C_B region at test, $F_{AB}(d_i)$ is the probability of d_i according to the cumulative density function estimated at training for the AB bound, and $F_{CD}(d_i)$ is the probability of d_i according to the cumulative density function estimated at training for the CD bound. Intuitively, the numerator corresponds to probability of responding B according to the AB bound, and the denominator normalizes according to the probability of responding C according to the CD bound. This extra step is necessary because the AB bound does not allow for calculating the probability of responding C, and the CD bound does not allow for calculating the probability of responding B. The probability of identifying a stimulus as being located in the C_C region at test is simply $p(C_C|d_i) = 1 - p(C_B|d_i)$.

Decision function

Both the density and boundary models used a common decision function that is described by:

$$p(R_X|d_i) = \frac{e^{\alpha p(C_X|d_i)}}{e^{\alpha p(C_X|d_i)} + e^{\alpha p(C_Y|d_i)}} \tag{A4}$$

where $p(R_X|d_i)$ is the probability of responding category X (denoted R_X) when stimulus d_i is present, $p(C_X|d_i)$ is the probability of locating stimulus d_i in the C_X region (as calculated by Eq. A1, Eq. A2, or Eq. A3), and α is a noise parameter estimated by minimizing the sum of square errors (SSE). The probability of responding category R_Y is simply given by $p(R_Y|d_i) = 1 - p(R_X|d_i)$.

Model selection

It should be noted that both models can predict perfect accuracy on the training data (Hélie, Shamloo, & Ell, 2017). However, the model predictions differ drastically at test. Because the boundary model is a special case of the density model, and the density model has twice as many free parameters, the density model is guaranteed to always fit the training data at least as well as the boundary model. To avoid this problem, model selection is performed by estimating the generalization error of the models on the test data using cross–validation (Busemeyer & Wang, 2000; Hélie, 2006).

The following procedure was repeated separately for each participant. First, each model (i.e., Eqs. A1, A2, and A4) was fit to the data from the last 200 training trials (Blocks 4–5). These trials were selected because performance is more stable at the end of training (see Figure 3). Next, the generalization error was calculated on the test data (Block 6), without refitting the model parameters (using Eqs. A1, A3, and A4) (Hastie, Tibshirani, & Friedman, 2001). If the density model had the smallest generalization error, then it was inferred that the participant learned a within–category representation. If the boundary model had the smallest generalization error, then it was inferred that the participant learned a between–category representation. This fitting procedure was applied individually to each participant in each condition.