The Evaluative Role of Rostrolateral Prefrontal Cortex in Rule-Based Category Learning

Dmitrii Paniukov & Tyler Davis

Department of Psychological Sciences, Texas Tech University

Address:

Texas Tech University, Department of Psychological Sciences, Box 42051 Lubbock, TX, 79409-2051, United States of America

Email addresses of authors: dmitrii.paniukov@gmail.com, tyler.h.davis@ttu.edu

Corresponding author is Dmitrii Paniukov. Preferred method of communication is email: dmitrii.paniukov@gmail.com

Abstract

Category learning is a critical neurobiological function that allows organisms to simplify a complex world. Rostrolateral prefrontal cortex (rlPFC) is often active in neurobiological studies of category learning. rlPFC has been associated in the previous category learning literature with switching between rules and tasks that place higher demands on rule abstraction and evaluation. The goal of the present study was to clarify the role of rlPFC in category learning and dissociate switching and evaluation accounts. Two common types of category learning tasks, matching and classification, were utilized. The matching task involved matching a reference stimulus to one of four target stimuli. In the classification task, participants were shown a single stimulus and learned to classify it into one of two categories. Although nearly identical, matching and classification place differential demands on switching and evaluation. In matching, a rule can be known with certainty after a single correct answer. In classification, participants may need to evaluate evidence for a rule even after an initial correct response. This critical difference allows isolation of evaluative functions from switching functions. If the rlPFC is primarily involved in switching between representations, it should cease to be active once participants settle on a given rule in both tasks. If the rlPFC is involved in rule evaluation, its activation should persist in the classification task, but not matching. The results revealed that rlPFC activation persisted into correct trials in classification, but not matching, suggesting that it continues to be involved in the evaluations of evidence for a rule even after participants have arrived at the correct rule.

Keywords: category learning, rostrolateral prefrontal cortex, rule evaluation, rule

switching, fMRI

Highlights

- Differences between rule-based matching and classification tasks were highlighted.
- Rostrolateral prefrontal cortex is involved in evaluation of evidence for a rule in rule-based category learning tasks.
- Rostrolateral prefrontal cortex is functionally connected with the caudate nucleus, a region that may also play an evaluative role in rule learning.

1. Introduction

Friend or foe? Predatory or prey? Edible or poisonous? Category learning is a fundamental cognitive capacity that is critical for survival. Grouping objects into categories allows organisms to generalize information to novel examples and make inferences about future events. As a complex cognitive function, many brain regions are involved in category learning including the prefrontal cortex (PFC), medial temporal lobes, striatum, and visual cortex (Ashby & Maddox, 2005, 2011; Poldrack & Foerde, 2008; Seger & Miller, 2010; Smith & Grossman, 2008).

Although there are many types of category learning, one of the most studied types in cognitive neuroscience is rule-based category learning. In rule-based category learning, people learn a logical rule that can be used to determine whether items are members of the category or not. Many real world categories are associated with logical, albeit imperfect, rules. For example, members of the category mammals can often be categorized based on whether they have fur and nurse their young. Broadly speaking, rule-based categorization is thought to depend upon

executive cortico-striatal loops that connect the PFC and the head of the caudate nucleus (Alexander et al., 1986; Seger & Miller, 2010). Although early work with patients and fMRI tended to treat the PFC as being involved in executive functions as a whole (Konishi et al., 1998, Robinson et al., 1980), recent work in cognitive neuroscience has begun to test whether specific subregions of the PFC serve distinct mechanisms (Seger, 2008; Seger & Miller, 2010; Ma et al., 2016).

One critical PFC region that has thus far eluded a thorough explanation in terms of its role in category learning is the rostrolateral prefrontal cortex (rlPFC). The rlPFC is known to be involved in a broad array of higher-level cognitive functions including abstract symbolic reasoning and analogical problem solving (Green et al., 2006; Specht et al., 2009), relational category learning (Davis et al., 2017), and goal-directed reward learning (Spreng et al., 2010). It is often described as a seat of human reasoning powers as it is significantly larger in humans than other primates, and its development across childhood tracks the development of fluid reasoning capacities (Gogtay et al., 2004; Semendeferi et al., 2011).

In humans, increased rlPFC activation is often observed for rule-based tasks involving abstract symbolic (Specht et al., 2009) and/or relational reasoning (Davis et al., 2017; Gray et al., 2003; Wendelken & Bunge, 2009). The rlPFC's precise mechanistic role in these tasks has been described in a number of different ways. One characterization of the rlPFC focuses on switching between representations. For example, the rlPFC has been found to be more active when participants switch between cognitive sets in rule-based tasks (Konishi et al., 1998, 2002; Monchi et al., 2001; Strange et al., 2001; Liu et al., 2015), in tasks requiring switching between internally focused and externally focused attention (Burgess et al., 2007), and in reward learning

tasks when exploring the values of different choices (Daw et al., 2006).

Other studies of rlPFC function have focused on its role in forming abstract symbolic rule representations and testing and evaluating such rules (Badre & D'Esposito, 2009; Vendetti & Bunge, 2014; Wendelken et al., 2012). Classic examples of the rlPFC's involvement in symbolic rule use include its stronger engagement for rules requiring higher-order relational integration (rules that integrate over multiple lower-order relationships) in tasks akin to Raven's progressive matrices (Christoff et al., 2001; Kroger et al., 2002), and tasks requiring participants to answer whether such a higher order relation is present in a stimulus (Bunge et al., 2009; Nee et al., 2014). More recently, the rlPFC has also been found to be engaged in the incremental learning of rules across trials (Badre et al., 2010; Davis et al., 2017). In these cases, the rlPFC may not only be integrating relationships within stimuli to form rules, but also integrating or accumulating information across trials to evaluate evidence for a rule and test how it applies to new stimuli. Indeed, in Davis and colleagues (2017), the rlPFC was active early in learning as participants acquired a new relational rule, but then later only on trials in which they needed to apply the rule to novel stimuli. This evaluative role in integrating and accumulating information for a rule potentially connects the rlPFCs role in rule learning with recent findings of its engagement during meta-cognitive judgments (Fleming et al., 2012), and may explain why the rlPFC is often active during learning of rules that do not strictly involve higher-order relational integration within a stimulus or trial (e.g., Seger & Cincotta, 2006; Liu et al., 2015).

To test the distinction between rule-switching and evaluation accounts of rlPFC function, we compared activation during two commonly used rule-based tasks from the category learning literature — matching and classification tasks. Matching tasks, like the Wisconsin Card Sorting

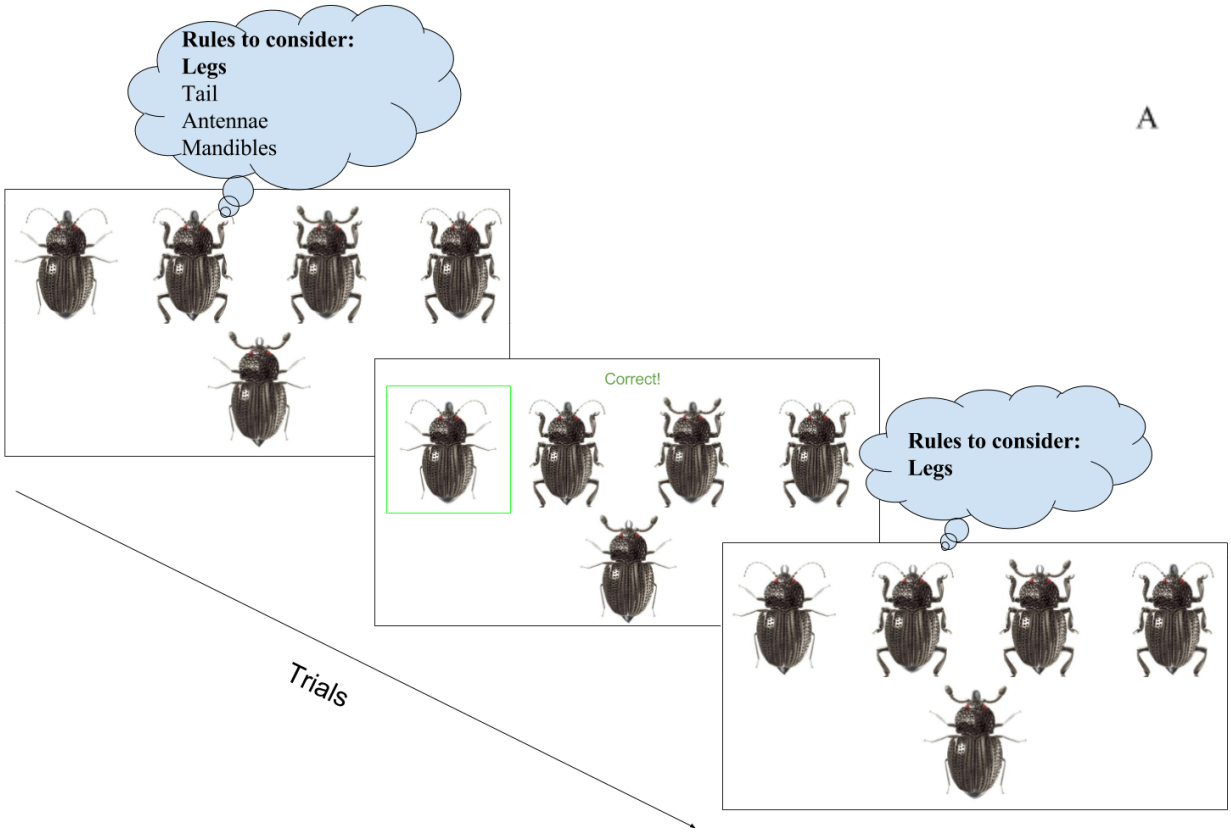
Test (Heaton, 1993), involve matching a multidimensional reference stimulus to a set of target stimuli that each match the reference on a single dimension. There is a rule that determines which target to match the reference to that is based on the stimulus dimensions. Participants learn the matching rule through trial and error. For example, in the matching task that we used in the present study, reference stimuli were schematic beetles that differed in terms of their legs, tail, antennae, and mandibles. Participants learned to match these reference stimuli to target beetles by choosing different candidate targets and receiving feedback. Often matching tasks will cycle through a number of rules forcing participants to abandon rules and shift to a new rule when old rules cease being productive.

As a neuropsychological measure, the primary process of interest in matching tasks like the Wisconsin Card Sorting Test is the process of shifting between cognitive sets to accommodate novel rules and suppress the previously correct rules. Consistent with the theory that the rlPFC governs representational switching, results from a number of neuroimaging studies have identified activation in rlPFC during trials in which the rule is switched (Konishi et al., 1998, 2002; Monchi et al., 2001; Strange et al., 2001; Liu et al., 2015). However, switch trials as well as the immediately following trials in which the novel rule is being acquired also place demands upon rule evaluation mechanisms. Participants must not only switch between the previous and new candidate rule representations, but also begin to test and evaluate evidence for new candidate rules. One critical aspect of simple matching tasks, like the Wisconsin Card Sorting Test, is that participants gain some conclusive information to evaluate a rule by on every trial: if a choice is wrong, a candidate rule can be eliminated; if a choice is correct, the rule is known with certainty. This aspect of matching tasks means that evaluation of evidence for a rule

and switching are strongly intertwined because participants will switch rules whenever they are wrong (Konishi et al., 1998; Monchi et al., 2001) and will quit evaluating new information once they are correct and proceed with applying the rule. Thus, for participants performing rationally, every possible switch trial is also a trial in which new information is being obtained and evaluated. Likewise, every trial that provides new information about the rule is a trial in which the participant has switched the rule they are using.

We contrast matching tasks with another common type of rule-based category learning — classification tasks. In classification tasks, participants learn a rule that allows them to sort multidimensional stimuli into two or more categories based on trial and error. When the rule is based on a single dimension, classification tasks involve very similar mechanisms to single dimension matching tasks like the Wisconsin Card Sorting Test. Participants test candidate rules, update their candidate rules based on feedback, and maintain rules in working memory once they have arrived at them. Indeed, because of this overlap in processing requirements, many neurobiological theories of category learning assume that matching and classification tasks have strong overlap in terms of the systems involved in acquiring and storing new rules and use results from both types of tasks interchangeably (e.g., Ashby & Maddox, 2005). However, classification tasks differ from matching tasks critically in requiring more extensive cross-trial evaluation of evidence for a rule, while otherwise keeping the demands on representational switching constant. In classification tasks, participants are shown single examples of stimuli and learn, using trial-and-error, to classify the stimuli into one or more categories based on the stimulus dimensions. Psychologically, as in matching tasks, participants must evaluate evidence for the correct rule by switching their attention between different stimulus dimensions to test

various candidate rules across trials. Also, like matching tasks, it is possible to eliminate a particular rule based on feedback in a single trial — if a participant tries the rule "thick legs = Category A", incorrect feedback will tell them that they need to switch the rule they are using and eliminate this rule. However, individual correct trials in classification tasks contain less information about the correct rule than in matching tasks. From an optimal observer standpoint, a single trial in a matching task can tell whether a rule is correct or incorrect, however, even when behaving optimally, on average, even several corrects trial in a classification task can leave potential candidate rules to decide amongst (see Figure 1 for an illustration of evaluative demands in the tasks). Interestingly, this greater demand for evaluating rules is not typically accompanied by an increase in switching. Extensive research suggests that participants in classification tasks use a win-stay-lose-shift strategy where they will not switch from a rule when they are getting correct feedback and will only switch to remaining candidate rules when they get negative feedback (Shepard et al., 1961; Nosofsky, Palmeri, and McKinley, 1993; Wilson & Niv, 2011; Niv et al., 2015). This critical asymmetry whereby correct trials in matching tasks indicate perfect knowledge of the rule whereas correct classification trials do not, suggests that comparing matching and classification may help to elucidate whether the rlPFC's role in category learning is more consistent with evaluation or switching accounts.


Rules to consider:
Legs
Tail
Antennae
Mandibles
Correct!
Rules to consider:
Legs
Trials

A

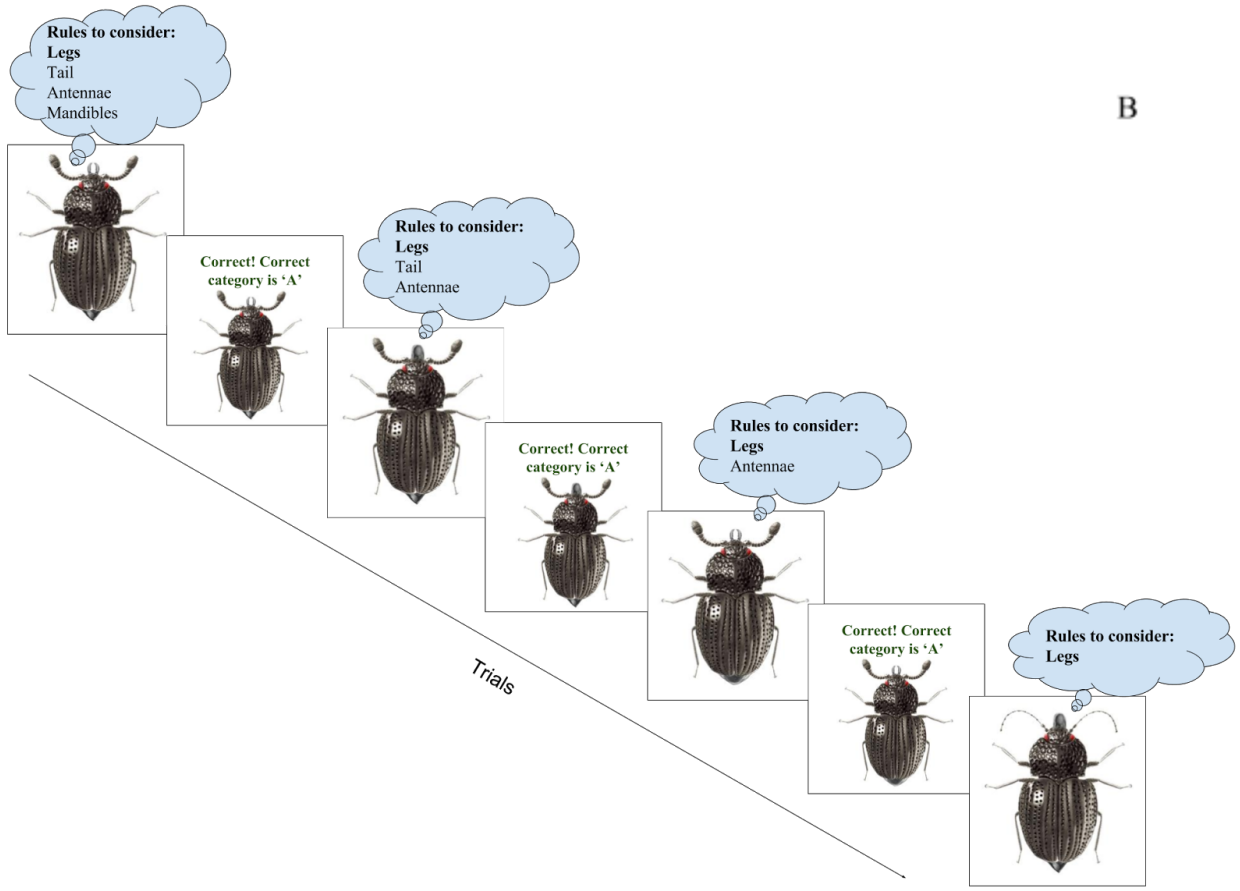


Figure 1. Examples of matching (A) and classification tasks (B) and how rules are eliminated in both tasks. In matching tasks, when the rule is correctly selected, all other rules are eliminated. Contrastingly, in classification tasks, even if a correct rule is selected initially, from an optimal observer's standpoint, several additional trials may be necessary to rule out other rules that are possible given the history of stimulus-category pairings participants have seen up to that point. In both of these cases, participants have arrived at the correct rule dimension ("legs" in bold) on the first trial of the depicted sequence. In matching, a single instance of correct feedback establishes that "legs" are the rule-dimension. In classification, although the participant may start out using the legs rule, they may only fully eliminate other possible rules that are consistent with the

stimulus history after a number of correct trials. For example, if in the first trial the participant chooses category 'A' for the thin legged beetle and gets correct feedback, the rules "Pointy tail = 'A'; Thick antennae = 'A'; and bisected mandibles = 'A'" are all still active. They only become eliminated after the participants encounters variations of the "thin legs = 'A'" rule, in which these candidate rules do not also hold.

To test how differences in evaluation demands between matching and classification tasks and how they impact rlPFC activation, we had participants complete the matching and classification tasks using schematic beetle stimuli. In both tasks, participants would iteratively learn rules via trial and error. Once rules were learned (defined by four correct trials in a row) the rule would switch and the participant would begin learning a new rule. For each rule there was therefore a *rule learning* phase, when participants were narrowing down and switching between rules, and a brief *rule application* phase, in which participants were using the final rule that they have arrived at (Figure 2).

	Rule learning				Rule application			
Trial #	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8
Feedback	Incorrect	Correct	Incorrect	Incorrect	Correct	Correct	Correct	Correct

Figure 2. An example of how the rule learning and rule application phases were separated for analysis. The last four or more contiguous correct trials were labeled as rule application. In this phase participants applied a final rule they arrived at. All prior to the last correct trials were labeled as rule learning. Participants were narrowing down and switching

between rules trying to find out the correct rule. All trials when participants did not respond were excluded from the analyses.

The different accounts of rlPFC function can be used to make predictions for the rule learning and application phases. If the rlPFC is primarily involved in switching, it should only be active during the rule learning phases of both matching and classification tasks, when participants are trying different rules and switching between them. Contrastingly, if the rlPFC is involved in evaluation, it should be active not only during the rule learning phase of both tasks, but also during the rule application phase for classification when participants have arrived at the correct rule, but are still evaluating evidence in support of it. Because rule evaluation is not necessary during matching, once the correct rule is found, rlPFC should not be active during the application phase of the matching task.

In addition to the rlPFC, we will also examine whether functional connectivity between the rlPFC and caudate nucleus is more consistent with the rule evaluation or switching account. Like rlPFC, the caudate nucleus had been implicated in rule evaluation and switching (Ashby & Maddox, 2005; Ashby et al., 1998; Filoteo et al., 2005; Liu et al., 2015; Monchi et al. 2001; Seger & Cincotta, 2006). Thus, to complement our main analyses, we used a Psychophysiological interaction (PPI) analysis that modeled connectivity between these regions during the rule-learning and application phases of the matching and classification trials. As with above, the rule evaluation account predicts that rlPFC and caudate connectivity should remain higher during the application phase of classification rules compared to matching.

To foreshadow our results, we found evidence consistent with the rule evaluation

hypothesis. The rlPFC activation during rule application for the classification task was significantly greater than rule application for the matching task. Greater rlPFC activation was also present for classification compared to matching during the rule learning phases. Likewise, there was significantly greater connectivity between an rlPFC seed and the caudate during both the rule application and learning phased for the classification task compared to matching. These results suggest that the rlPFC's role in category learning may be primarily rule evaluative and serve to accumulate and integrate evidence when acquiring novel rules.

2. Method

2.1. Participants

Twenty-seven participants were recruited from the Texas Tech University community via an electronically posted announcement. Participants were required to be at least eighteen years of age, right-handed, have a minimum of an eighth grade education, speak English fluently, and not have any contraindications for MRI research. Participants were compensated with $35 for the study. Two participants fell asleep during their scanning session, and therefore were removed from the further analyses. One participant opted out from the last two scanning runs, but the rest of his/her data were used in the analyses. The study was approved by Human Research Protection Program at Texas Tech University.

2.2. Stimuli and Procedure

Participants were asked to sign a consent form, MRI-safety checklist and complete a computer-based tutorial, where they were shown example trials and stimuli outside of the scanner. Upon completion of the screening forms and the tutorial, participants were placed into the scanner. In the scanner, participants completed four runs of the matching task and four runs

of the classification task in an order that was balanced across participants.

Two tasks, matching and classification, utilized sixteen images of schematic beetles. The beetles varied across four binary feature dimensions: legs, mandibles, antennae, and tail (see Figure 3 for two examples with opposite features). For both tasks, the sixteen stimuli were presented in blocks in which each stimulus was shown once in a random order.

Figure 3. Stimuli of schematic beetles with completely distinct feature dimensions. The feature dimensions were legs (thick or thin), mandibles (closed or bisected), antennae (fuzzy or dotted), and tail (pointy or round).

In the matching task, each screen contained a reference beetle and four target beetles, each of which matched the reference beetle on a single dimension (see Figure 4). The position of the beetles on the screen was randomized to minimize effects of feature salience and balance motor responses.

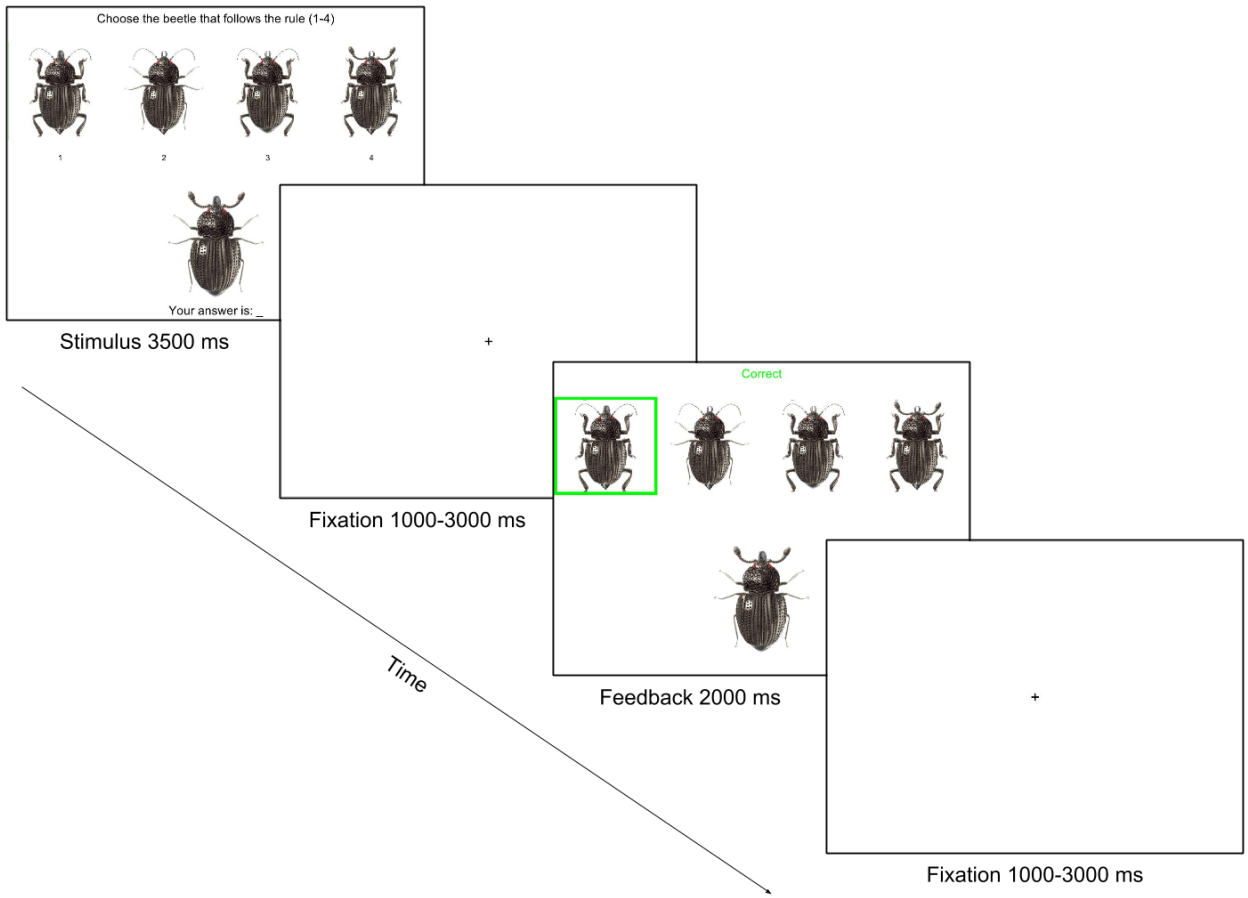


Figure 4. Example of the matching task. Participants saw a reference beetle on the bottom and four target beetles on top. The target beetles each matched the reference beetle on a single feature. Participants would select a target beetle then receive feedback about whether their choice was correct or incorrect.

In the classification task, each screen contained a single beetle (see Figure 5). The beetles

were assigned into category A or B, based on a random rule, defined by a single feature.

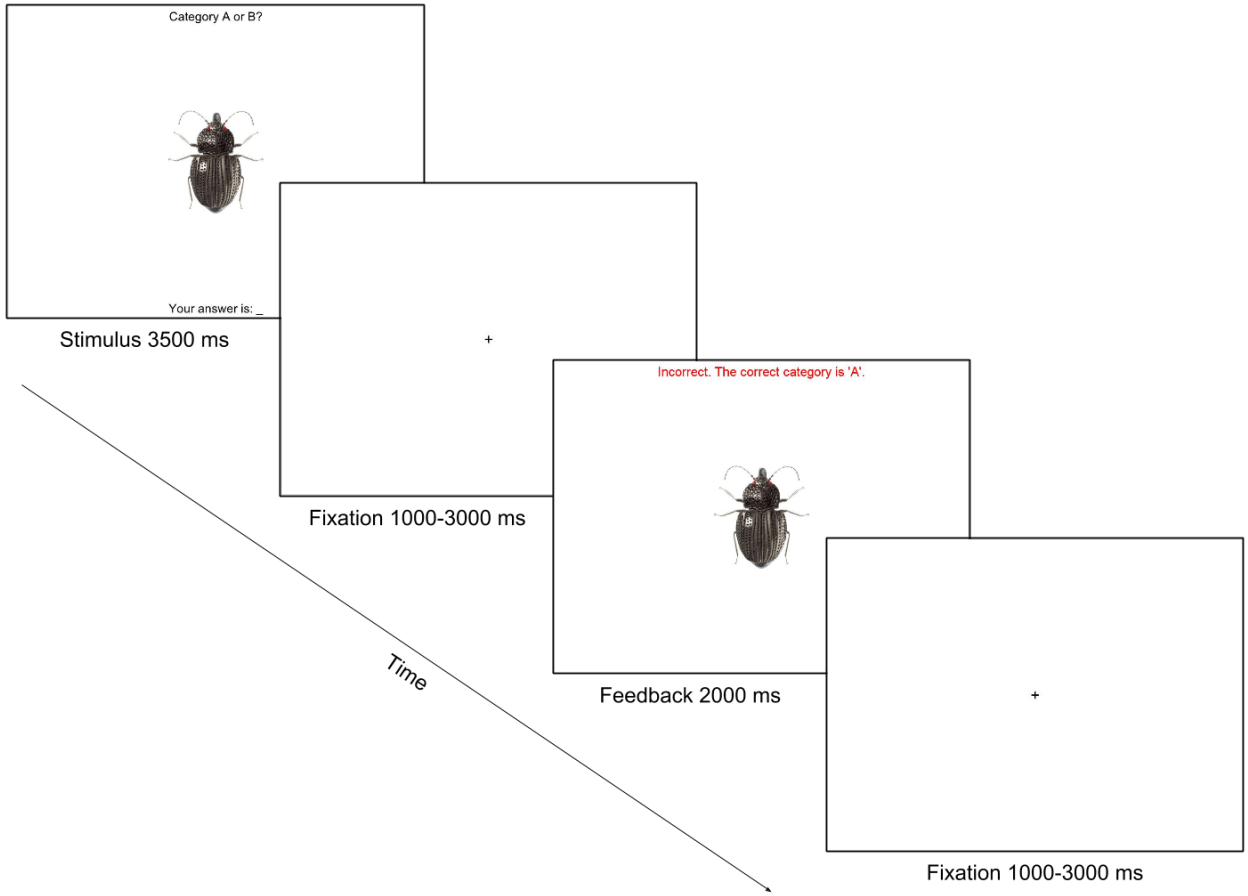


Figure 5. Example of the classification task. Participants saw a single beetle on the screen and were asked to categorize it as a member of Category A or B. After participants made the choice, they were given feedback about whether their choice was correct and the correct category label.

In both tasks, participants had 3.5 seconds to categorize a stimulus using a button box held in their right hand. After a brief fixation (1, 2, or 3s; mean = 2s), participants were provided with feedback for 2s, during which the beetle was presented again along with a message "Correct/Incorrect/Failed to respond" in the matching task or "Correct/Incorrect/Failed to respond." and "The correct category is A/B" in the classification task. After the feedback, a brief

fixation was presented (1, 2, or 3s; mean = 2s). Upon achieving a termination criterion of four correct answers in the row, the rule was switched to a new randomly selected rule. When a run was completed, participants were presented with a number of correctly solved rules during the run (i.e., "You successfully learned [number] categories! Good job!"). Participants had thirty-two trials in each run and eight runs total (four runs in a block for each task), followed by an anatomical scan. After the scan, participants were thanked, compensated and dismissed from the study.

2.3. fMRI Data Acquisition

The data were collected at Texas Tech Neuroimaging Institute using a Siemens Skyra 3T scanner with a 20-channel head coil. A T1-weighted sagittal MPRAGE was obtained with TR = 1.9s, TE = 2.49, flip angle = 9, matrix of 256x256, field of view = 240, slice thickness= 0.9 mm with a gap of 50%, 1 slice. T2-weighted BOLD echo planar images (EPI) were obtained with TR=2s, TE=30ms, flip angle = 90, matrix = 64x64, field of view = 192, slice thickness = 3x3x4 mm, 35 ascending axial slices, 156 volumes in each scanning run. The slice prescription was tilted off of parallel with AC/PC to reduce susceptibility artifact in orbital frontal cortex (Deichmann et al., 2003).

2.4. fMRI Data Preprocessing and Analysis

MRI data preprocessing included the following steps: DICOM imaging files were converted to NifTI files using dcm2nii from the Mricron software package (Rorden & Brett, 2000). Skulls were removed using the BET tool (Smith, 2002) from FSL software package (Jenkinson et al., 2012; Woolrich et al., 2009) for the BOLD EPI images and ANTs (Avants et al., 2009) with OASIS template (Avants & Tustison, 2014) for the T1 anatomical images.

Motion correction was carried out using a 6-DOF affine transformation implemented in FSL's MCFLIRT tool (Jenkinson et al., 2002). Data were smoothed with an 8 mm FWHM Gaussian kernel and highpass filtered (100s cut off). Finally, data were manually checked for quality issues such as visual spikes produced by the scanner, incorrect brain extraction, and excessive motion. The quality check revealed visual artifacts in the first run for four participants and thus these runs were excluded from further analysis.

Functional MRI data analysis was carried out using a standard three-level analysis in FSL's FEAT, as implemented in Nipype (Gorgolewski et al., 2011). The first-level analysis consisted of prewhitening using FSL's FILM (Woolrich et al., 2001) to account for temporal autocorrelation, task based-regressors (see below) convolved with double-gamma hemodynamic response, and their temporal derivatives. Additional confound regressors of no interest included six motion parameters, their temporal derivatives and regressors for scrubbing high motion volumes exceeding a framewise displacement of 0.9mm (Siegel, et al., 2014). The same high pass filter setting that was used to process the fMRI data was used to process the design matrix. First-level statistical maps were registered to a standard space in a two-stage registration consisting of (1) registration of each BOLD timeseries to respective participants' T1 MPRAGE using the BBR algorithm (Greve & Fischl, 2009) and (2) registration of the T1 to the standard space (MNI-152 brain template) using nonlinear ANTs registration (Avants et al., 2009). Second-level analysis combined across runs within a participant, and was carried out using a fixed effects model in FLAME (Beckmann et al., 2003; Woolrich et al., 2004; Woolrich, 2008). Second-level regressors included task type (matching vs. classification). Third-level analyses examined the effects of first and second-level contrasts across participants. Final thresholding of

statistical maps at $p < 0.05$ was done via a cluster mass correction (with primary/cluster-forming threshold of $t(24)=2.49$) (Bullmore et al., 2000) for multiple comparisons implemented in FSL's non-parametric permutation tool Randomise (Winkler et al., 2014) with variance smoothing set at 8 mm.

All trials in both tasks were split into rule learning phase in which participants were acquiring the correct rule, and a rule application phase in which participants had acquired the rule and were now applying it (see Figure 2 for an example of the phases for the classification task). In the matching task, all correct trials were considered as rule application trials due to the nature of the task, and all incorrect trials as rule learning. In the classification task only the last four or more continuous correct trials were considered as rule application, and the rest of the trials as rule learning.

Two univariate models were used to model the fMRI data. Model 1 used six explanatory variables (EVs): (1) stimulus presentation (onsets) for rule application trials, (2) onsets for rule learning trials, (3) onsets for trials when participants did not answer, (4) correct feedback, (5) incorrect feedback, and (6) feedback for trials when participants did not answer. These EVs were compared in three contrasts: (1) rule application > baseline, (2) rule learning > baseline, (3) rule application > rule learning. Statistical maps describing the results of these contrasts were thresholded to correct for multiple comparisons at the whole-brain level and within an a priori frontal pole mask (ROI), which was defined as the max-probability frontal pole region from the Harvard-Oxford atlas included in FSLView (Desikan, et al., 2006).

A second model (Model 2) tested whether the task variables modulated the connectivity between rlPFC and other brain regions using a psychophysiological interaction (PPI) analysis.

Rostrolateral PFC timeseries signal was extracted from an 8mm sphere around the highest peak found in the rule learning > baseline contrast for classification trials (MNI coordinates: x=-40, y=54, z=4 mm). The EVs in Model 2 included all six EVs from Model 1 in addition to three new EVs, namely (7) the rlPFC timeseries, (8) the interaction between EV 1 (onsets for rule application) and EV 7, and (9) interaction between EV 2 (onsets for rule learning) and EV 7. No convolution, temporal filtering or temporal derivatives were used for the rlPFC timeseries EV and its interactions. Three contrasts examined the (1) interaction of rule application onsets with rlPFC timeseries vs. baseline, (2) interaction of rule learning onsets with rlPFC timeseries vs. baseline, and (3) interaction of rule application onsets with rlPFC timeseries vs. interaction of rule learning onsets with rlPFC timeseries. The Harvard-Oxford max-probability caudate nucleus mask was used as an a-priori ROI in Model 2.

3. Results

3.1. Behavioral results

For the behavioral results we used a Poisson mixed effects model from the LME4 package (Bates et al., 2015) and a mixed effects model Cox regression survival model from the Coxme package (Therneau, 2015) in R (R Core Team, 2014) to examine how number of rules solved and trials to the rule termination criterion differed between task types. Consistent with the assumption that the classification task placed more demands on rule evaluation over trials than matching, participants solved more rules during matching (M=15.6, SD=4.14) than during the classification task (M=9.92, SD=3.22), z=5.58, p<.001. Likewise, participants took fewer trials to reach a rule termination criterion in matching (M=6.56, SD=4.67) than in the classification task (M=11.73, SD=7.56), z=12.11, p<.001.

3.2. Imaging results

To examine the neuroimaging results, all trials were sorted into either rule learning or rule application phases. In the matching task, all correct trials were considered as rule application trials due to the nature of the task, and all incorrect trials -- as rule learning. In the classification task only last four or more contiguous correct trials were considered as rule application, and the rest of the trials as rule learning. Trials when participants failed to respond were not included into any phase.

3.2.1. Rule learning and rule application in the matching task.

Both the switching and rule evaluation accounts of rlPFC function predict that there should be more rlPFC activation during the learning phase of the matching task. Rule learning should require more evaluation demands than application because rules are known with full certainty once a correct answer is found, and thus participants will likely cease intentionally evaluating additional evidence for a rule. Likewise, prior to arriving at the correct rule, participants should switch between rules when they are presented with incorrect feedback. Consistent with these accounts, we found a cluster (659 voxels) in the rlPFC ROI that activated greater for learning than application trials.

Additional areas revealed to be more active during learning than application in the matching task were parietal and lateral occipital regions. These results are consistent with other studies examining activation during learning phases of similar rule-based category learning tasks (Seger & Cincotta, 2006; Liu et al., 2015), and are consistent with previous suggestions that the superior parietal cortex is involved with cognitive control functions during rule-based category learning. Whole-brain results for the opposite contrasts (rule application > rule learning) were

also consistent with previous findings on rule-based learning (Seger & Cincotta, 2006; Nomura et al., 2007) with greater activation in the MTL (hippocampus and parahippocampal gyrus) during rule application. Theories of category learning suggest that the MTL is involved in the long-term retention and retrieval of category information during category learning (Ashby & Maddox, 2011; Ashby et al., 2011; Davis et al., 2012a; 2012b). Additional regions more active for rule application included the ventromedial prefrontal cortex, a region we have recently identified as being associated with stronger decision evidence during categorization (Davis et al. 2017), putamen, parietal cortex, occipital cortex, temporal cortex, and insular cortex (all brain maps and other project details are posted at https://osf.io/ge8vf/).

3.2.2. Rule learning and rule application in the classification task.

Similar to the matching task, parietal and occipital regions were activated in the rule learning > rule application contrast for the classification task. Additional areas that were more active for rule learning than application included the dorsolateral PFC, which is thought to be involved in maintaining and manipulating information in working memory during rule-based categorization (Monchi et al., 2001; Filoteo et al., 2005; Seger & Cincotta, 2006), and the cerebellum. No regions were significantly more active in rule application compared to rule learning (rule application > rule learning) in the whole brain results.

Our primary region of interest, rlPFC, was significantly activated relative to baseline in both rule learning > baseline and rule application > baseline contrasts in both the whole brain and ROI analyses (see Figure 6). Further, there was no difference between rlPFC activation during rule learning and application for the classification task within either the ROI or whole brain analysis, even at liberal uncorrected thresholds. Together, these results suggest that the

rlPFC is engaged in a more rule evaluative role that persists into the application phase in the classification task. This result is inconsistent with switching accounts because participants generally would not be switching rules once they have arrived at a rule that does not result in negative feedback. However, they may continue evaluating confirmatory evidence for their chosen rule because the rule is rarely known with certainty after the first correct trial in the classification phase.

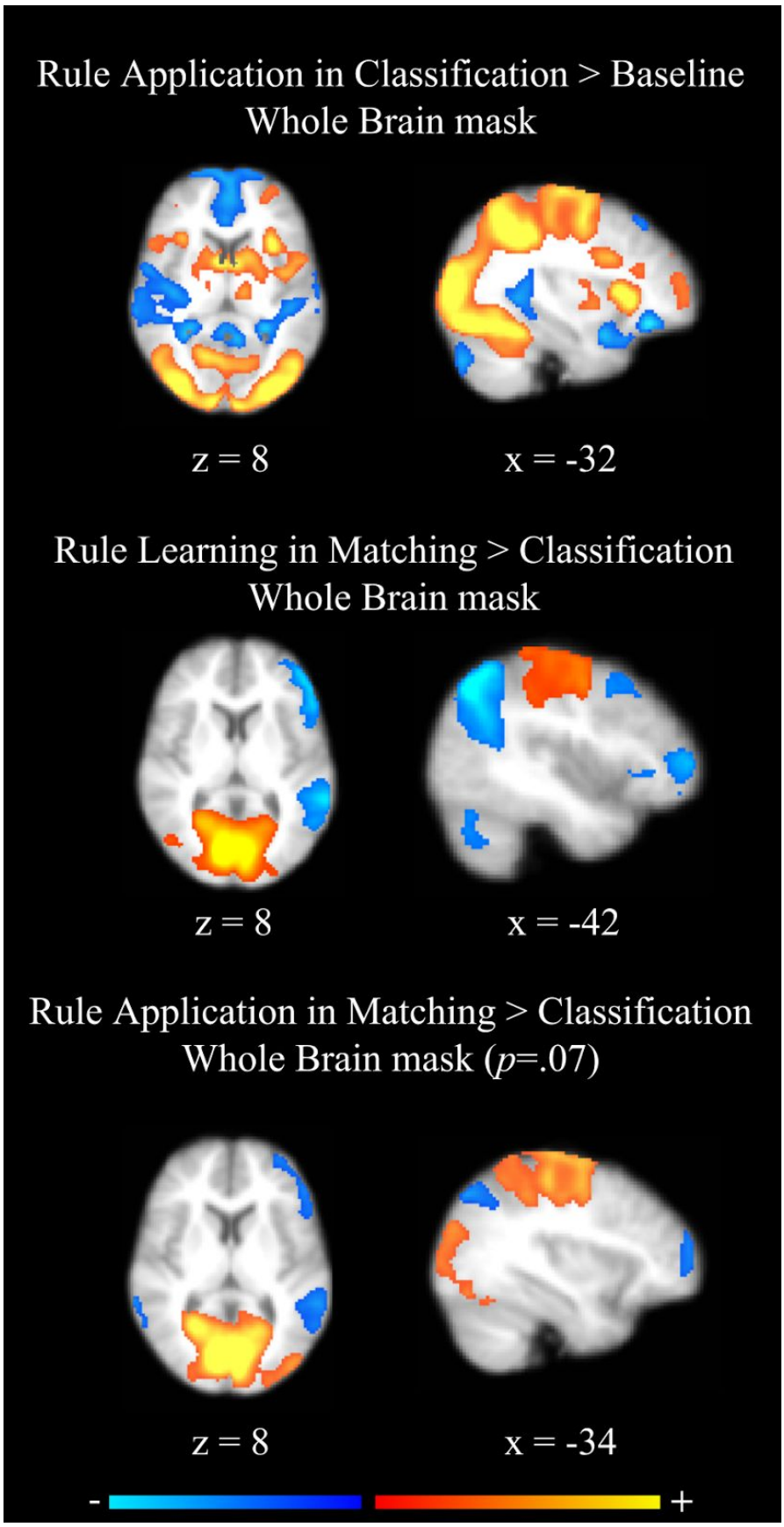

Rule Application in Classification > Baseline
Whole Brain mask
z = 8
x = -32
Rule Learning in Matching > Classification
Whole Brain mask
z = 8
x = -42
Rule Application in Matching > Classification
Whole Brain mask (p=.07)
z = 8
x = -34
-
+

Figure 6. Brain activations for the main contrasts. Matching > classification contrasts are in red-yellow and classification > matching in blue. The top contrast shows that the rlPFC was more active in the classification task during the rule application phase. The middle contrast shows that the rlPFC was more activated in the classification vs. matching task in the rule learning phase. These contrasts along with the ROI contrasts show that the rlPFC was involved within and between the tasks. The contrast on the bottom answers the main question of the study: the rlPFC was more active in the classification task than in matching for the rule application, thus suggesting that rlPFC was associated with rule evaluation.

3.2.3. Comparing rule learning in the matching and classification tasks.

Occipital and parietal cortices as well as thalamus were more active during the rule learning phase for matching than for classification. The classification task activated more rlPFC (in both whole brain and ROI analyses), in addition to occipital, parietal, temporal, lateral prefrontal cortices and cerebellum. The parietal cortex was previously associated as a region working with rlPFC on relational integration (Wendelken et al., 2012). According to the hierarchical control theory (Badre & D'Esposito, 2007), the lateral PFC regions, more posterior to the rlPFC, could be involved in learning of less complex rule information.

3.2.4. Comparing rule application in the matching and classification tasks.

The primary goal of the current study was to compare rule evaluation versus switching accounts of rlPFC function. Because participants, by definition, quit switching rules during rule application in both matching and classification tasks, but the classification task often involves continued uncertainty and evaluation of evidence for a rule, contrasting rule application during matching and classification allows us to isolate rule evaluation mechanisms. Consistent with the

evaluation account, rlPFC was more active in the classification task compared to the matching task. This difference was significant for the a priori ROI analysis and marginal in the whole-brain analysis ($p = 0.07$). In addition to rlPFC, there was greater activation in superior parietal, temporal, and lateral occipital regions, in the classification task, and inferior parietal and medial occipital regions in the matching task (see Table 1). The superior parietal regions have been associated with measures of relational integration that is also one of the functions associated with rlPFC (Vendetti & Bunge, 2014; Wendelken et al., 2012), and with measures of decisional uncertainty in the broader categorization literature (Davis et al., 2017; Grinband et al. 2006; Degutis & D'Esposito 2007; Braunlich and Seger 2016).

Table 1

Rule application in the matching and classification tasks

Cluster Region(s)	**Cluster Size (voxels), Cluster p-value**	**Local Maximum Region**	**t-value**	**MNI coordinates (x,y,z)**		
Matching > Classification						
Occipital Cortex Superior Parietal Cortex Cerebellum	29803, p=0.0002	Lingual Gyrus	12.7	12	-74	-4
		Intracalcarine Cortex	12.4	10	-80	4
		Lingual Gyrus	11.9	-12	-76	-10
		Intracalcarine Cortex	11.2	-8	-84	2
Classification > Matching						
Occipital Cortex Posterior Parietal Cortex Temporal Cortex	3768, p=0.017	Supramarginal Gyrus, posterior division	6.22	-48	-50	56
		Middle Temporal Gyrus, posterior division	5.69	-62	-28	-4
		Lateral Occipital Cortex, superior division	5.27	-42	-68	48
		Angular Gyrus	4.87	-60	-58	12

Occipital Cortex Posterior Parietal Cortex Temporal Cortex	2033, p=0.0366	Angular Gyrus	5.66	54	-50	50
		Angular Gyrus	5.42	60	-48	38
		Middle Temporal Gyrus, temporooccipital part	5.19	66	-46	-4
		Lateral Occipital Cortex, superior division	4.53	52	-62	40

Note: Local maximum regions are based on the Harvard-Oxford atlas. The coordinates are in the standard MNI space. The p-values are taken from non-parametric tests produced by Randomise tool in FSL.

3.2.5. Connectivity analysis.

Although our primary hypotheses about the rlPFC were about its activation during rule application for classification and matching, it is also useful to examine connectivity between the rlPFC and other regions. One a priori region we were interested in due to its prominence in theories of rule-based categorization is the caudate nucleus. The caudate nucleus is another brain region that is known to be involved in rule learning and has been ascribed a number of different roles including hypothesis testing and rule switching. (Ashby & Maddox, 2005; Ashby et al., 1998; Filoteo et al., 2005; Liu et al., 2015; Monchi et al. 2001; Seger & Cincotta, 2006). Hypothesis testing itself is a complex process that involves both testing of different rules and evaluation of evidence for a rule. In the broader literature, the caudate is involved with many tasks, including procedural task chunking and habit formation (Graybiel, 2008), so one possibility is that it interacts with the rlPFC to facilitate rule evaluation during learning. To test this hypothesis, we ran a psychophysiological interaction analysis (PPI) using rlPFC as a seed region. Consistent with the idea that the caudate interacts with rlPFC to facilitate rule evaluation, we found stronger connectivity between the rlPFC and caudate during both rule learning and application for the classification task compared to the matching task. This clarifies the role of the

caudate in hypothesis testing as being involved specifically with the evaluation of evidence for a rule.

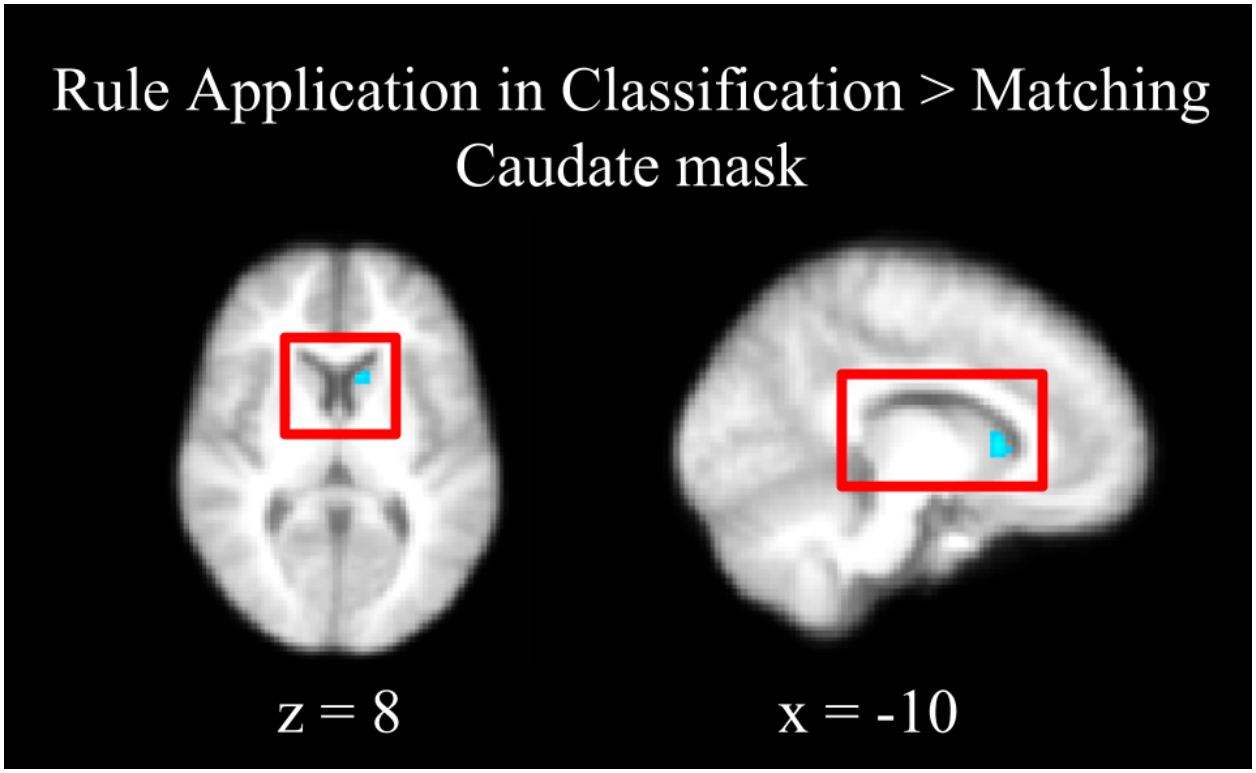


Figure 7. The caudate nucleus was active in rule application for classification vs. matching contrast. The result is constrained to an anatomically defined caudate mask. The box signifies that the results are not for the whole brain.

4. Discussion

The goal of the current study was to compare rule evaluation and switching accounts of rlPFC function in rule-based category learning. The rlPFC has been established as a critical brain region for many higher-order cognitive capacities, yet so far theories of category learning have not fully established a role for the rlPFC. Based on the broader literature, we developed two contrasting accounts suggesting that the rlPFC is involved in rule switching or in the evaluation

of evidence for a rule. We tested these contrasting accounts by comparing rlPFC activation during rule-based categorization tasks requiring either matching or classification learning. Although often treated as the same, these two tasks place different demands on rule evaluation mechanisms. In matching tasks, a rule can be known with certainty after the participant answers a single question correctly and can be eliminated with certainty with every incorrect answer. Therefore participants will tend to evaluate evidence for a rule on the same trials in which they switch between rules. This is not the case in classification tasks, where eliminating a rule can be accomplished in a single trial, but it may be necessary to evaluate evidence over several correct trials before a rule can be established with full certainty. Given this asymmetry in switching and evaluation between matching and classification tasks, we were able to isolate evaluation mechanisms by comparing activation on trials in which participants were applying a rule in matching and classification tasks. Consistent with the hypothesis that the rlPFC is involved in evaluation of evidence for a rule, we found that the rlPFC remained active during rule application in classification learning, but was not active during rule application in matching.

The current results are consistent with other recent studies revealing rlPFC involvement in category learning. Seger & Cincotta (2006) and Liu et al. (2015) found the rlPFC was more active during rule learning compared to rule application in classification-like tasks, and suggested the rlPFC is a part of a “cognitive” cortico-striatal loop involved in rule-based categorization. The current findings are consistent with this hypothesis but go further in suggesting that there may be critical differences between rule-based learning tasks in terms of the demands they place on these rule evaluation mechanisms. This is an important finding as results from matching and classification tasks have often been used interchangeably in the literature on

the neural basis of categorization (e.g., Ashby & Maddox, 2005). Our findings suggest that some fundamental contrasts (e.g., rule application vs. learning) can differ markedly between the types of rule-based task employed.

By revealing critical differences between types of rule-based tasks, our results fit well within the recent multiple systems literature that examines whether there are differences in the neural systems recruited between types of categorization tasks (rule-based vs. procedural learning; Ell et al., 2006; Nomura et al., 2007; See Ashby & Ell, 2001; Ashby & Maddox, 2005; Schnyer et al., 2009; Seger and Miller, 2010, for reviews) or even amongst tasks of the same basic type (e.g., multiple prototype learning systems; Ziethamova et al., 2008). Although it is important to note here that we are not suggesting that matching and classification involve wholly separate systems, only that they differentially load on these systems, leading to system-level differences in BOLD activation. Future research attempting to build a comprehensive account of the neural basis of categorization will need to take more care in considering commonalities and differences between types of rule-based task. Consideration of the demands of particular rule-based tasks may also be critical for neuropsychological assessment where matching tasks, like the Wisconsin Card Sorting Test, are currently the most popular type of task, but may not be as diagnostic of frontal rule evaluation processes as classification learning.

The present results may also inspire future research on rlPFC function in animal models. Recent work on rlPFC function in animal models has demonstrated a critical role for the rlPFC in rapid, one-shot learning of rules, but not in application of well-learned rules (Boschin et al., 2015). Our matching results are consistent with this role of rapid, one-shot rule learning in that the rlPFC was engaged up until participants learned the rule, but then activation dropped to

baseline during even early application. However, the classification results suggest that the rlPFC may be engaged even as participants begin to apply the rule, if they need to evaluate additional evidence for the correct rule (e.g., rule out remaining alternatives). Putting these results together suggests that rlPFC can allow one-shot learning, but whether it is engaged for longer-term rule acquisition depends on the demands of the task. Future work with animal models on the rlPFC function would benefit from examining tasks, like our classification task, where evidence for a rule must be accumulated and integrated across trials.

Comparing to Boschin et al.'s (2015) results with rule-learning in macaques, one limitation of our study is that we did not have any extended application trials in which participants were applying very-well learned rules for which they had already achieved automaticity. Thus it is not possible, within the current data, to establish whether the rlPFC would remain more engaged for classification learning relative to matching, or if eventually rlPFC involvement would decrease. Given the rapid shift in rlPFC involvement during matching, where it was active during learning but declined in activation as soon as participants began applying the rule, it is likely that rlPFC activation would also decline rather rapidly in classification once the rule was known with certainty. This very brief role for the rlPFC would be consistent with Boschin's findings from rule learning in macaques as well as other recent studies examining rlPFC involvement in category learning. For example, we recently found, in a relational category learning task (Davis et al., 2017), that rlPFC activation was high early in learning, but in later test phases was only re-engaged when participants needed to generalize the rule to novel relational examples, not for applying the rule to well-learned examples. In the future, it will be important to test whether the rlPFC exhibits similar trajectories in basic

rule-based classification tasks.

Outside of the immediate importance of this work for research on rule-based category learning, the present study adds to a growing literature on rlPFC function in higher-level cognition. Just as it has in category learning, ascribing a single cognitive function to rlPFC has been difficult due to its activation in a wide range of tasks (Duncan & Owen, 2000; Gilbert et al., 2006). For example, the rlPFC has been found to not only be involved in switching (Konishi et al., 1998, 2002; Monchi et al., 2001; Strange et al., 2001) and abstract rule evaluation (Christoff et al., 2001; Kroger et al., 2002; Vendetti & Bunge, 2014; Wendelken et al., 2012; Bunge et al., 2009; Davis et al., 2017; Nee et al., 2014), but also in reinforcement learning (Daw, et al., 2006) and metacognition (Fleming et al., 2012; 2014). Perhaps the best candidate for uniting across these disparate areas of cognition comes from the hierarchical control literature, which suggests that that the rlPFC sits atop a rostro-caudal gradient of rule abstraction (Badre, & D'Esposito, 2007; 2009, Badre et al., 2010).

In much of the previous research on hierarchical control theory, participants are given the rules to guide their behavior (e.g., Badre & D'Esposito, 2007) and/or they learn them in single trials (e.g., in Raven's-like tasks). In these cases, the minimum abstractness or rule complexity needed to follow or solve the rule is related to which part of the PFC is engaged, with the most abstract rules being processed in rlPFC. Under hierarchical control theory, it is somewhat surprising that rules like those used in our tasks, or metacognitive evaluations of purely perceptual processes (Fleming et al., 2012; 2014), would require the use of the rlPFC. One possibility to reconcile these differences within hierarchical control theory is that the rlPFC is involved whenever participants use rules that involve a structured predicate-argument

representation (Ramnani & Owen, 2004; Vendetti & Bunge, 2014), even when these may not be strictly necessary in the task. In the present task, participants may use such representations to guide their initial rule learning and evaluation, but then move back along the control hierarchy to more perceptually based representations after the rules are fully learned and established. Under this hypothesis, the rlPFC may be necessary to learn certain abstract rules, but may be engaged for learning any rule depending on how participants approach it in a given context. Rules based on abstract relations may always depend on rlPFC, whereas rules based on very elementary perceptual features, like the Gabor patches used in Davis et al. (2017) may seldom recruit rlPFC. Between these two extremes may be rules like those in the present task, which can be solved using symbolic representations or perceptual representations (or a combination of both). Whether participants use a symbolic or more perceptual strategy may depend on how easily relevant differences between stimuli are encodable symbolically (e.g., using language; Davis et al., 2009). Although this explanation for integrating the present results with hierarchical control theory is plausible given the broader literature on rlPFC function (Ramnani & Owen, 2004; Vendetti & Bunge, 2014), it remains to be seen whether it can account for rlPFC's role in other primarily evaluative domains such as in metacognitive judgments of perception (Fleming et al., 2012; 2014).

Relative to some tests of hierarchical control theory, our study had several limitations in terms of differences in the display and number of options available during the matching and classification tasks. The matching task had four options available (match to the beetle with the same legs, mandibles, antennae, or tail) on every trial whereas in the classification task only two options were available (choose category A or B). Likewise, the matching task had more visual

clutter on the screen with the four target beetles, whereas in classification a single beetle was presented. Although we observed various brain regions associated with visual, motor, and feedback differences between the tasks, the main region of interest, the rlPFC, was found to track predictions from the stimulus evaluation hypothesis both within and between tasks. For example, consistent with the prediction that the rlPFC would not be engaged in the matching task once the rule was known with certainty, rlPFC was engaged during learning but not during rule application in the matching task. Relatedly, consistent with the idea that rule evaluation would persist into the rule application phase in classification, the rlPFC was engaged relative to baseline during both learning and application in the classification task. Because demands were equivalent within task, this pattern where the difference between rule-learning and application was greater in matching than classification cannot be driven solely by button or stimulus display differences. Finally, although theory does not suggest rlPFC should be sensitive to differences in stimulus display or response options, if the rlPFC were sensitive solely to task complexity, the fact that the matching task is more complex in both respects should work against our hypothesis that it would be more engaged during classification.

Another methodological question that arises from the present results is whether our results would hold across all matching and classification tasks. In the current study, the matching and categorization tasks both rely on rules based on a single dimension (i.e., unidimensional rules), which is partially responsible for creating the asymmetry between switching and evaluation demands. Matching tasks can involve more complex conjunctive rules that combine across multiple features (e.g., match the beetle based on legs and antennae). In such cases, depending on the design, it may be possible to have correct trials prior to when the full rule is

known with certainty, just like in classification. Thus it may be possible to examine differences between rule-evaluation and switching mechanisms using only matching-based tasks. For example, one could compare rule application using conjunctive and unidimensional rules. In the present case, we opted instead to keep rule complexity constant and compare across task types, which also gave us the added benefit of being the first to directly compare these two tasks that have largely been assumed to be interchangeable in the categorization literature. In the future, it may be useful to test whether our results extend to comparing tasks of the same type (e.g., matching) that vary in rule complexity.

In conclusion, the present study examines the role of the rlPFC in category learning and differentiates between two accounts of rlPFC function that have been discussed in this literature, rule switching and rule evaluation. To test these accounts, we compared activation during different phases of two types of rule-based tasks, matching and categorization, that differ in their demands on evaluation. Consistent with the evaluation hypothesis, the rlPFC was active for tasks requiring more evaluation demands, even when participants were not switching between rules. These results are critical because they help to establish a role for the rlPFC in category learning literature and because they highlight novel differences in the systems engaged for two rule-based tasks that have been used interchangeably in much of the neurobiological literature on category learning thus far. Future research can build on these results by investigating how different types of rule-based representations and task demands impact involvement of rlPFC and its connection with memory systems involved with rule learning, such as the striatum.

5. References

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*(1), 357-381.

Ashby, F. G., Alfonso-Reese, L. A., Turken, 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. (2001). The neurobiology of human category learning. *Trends in cognitive sciences*, *5*(5), 204-210.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147-161.

Ashby, F.G., Paul, E.J., & Maddox, W.T. (2011). COVIS. In E. M. Pothos & A.J. Wills (Eds.), *Formal approaches in categorization*, 65-87. New York: Cambridge University Press.

Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS). *Insight Journal*, *2*, 1-35.

Avants, B. B.& Tustison, N. (2014). *ANTs/ANTsR Brain Templates*. Retrieved: 13 34, Dec 27, 2016 (GMT) from: https://dx.doi.org/10.6084/m9.figshare.915436.v1

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315-326.

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of cognitive neuroscience, 19*(12), 2082-2099.

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical?. *Nature Reviews Neuroscience*, *10*(9), 659-669.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage*, *20*(2), 1052-1063.

Boschin, E. A., Piekema, C., & Buckley, M. J. (2015). Essential functions of primate frontopolar cortex in cognition. *Proceedings of the National Academy of Sciences*, *112*(9), E1020-E1027.

Braunlich, K., & Seger, C. A. (2016). Categorical evidence, confidence, and urgency during probabilistic categorization. *NeuroImage*, *125*, 941-952.

Bullmore, E., Horwitz, B., Honey, G., Brammer, M., Williams, S., & Sharma, T. (2000). How good is good enough in path analysis of fMRI data? *NeuroImage*, *11*(4), 289-301.

Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *Neuroimage*, *46*(1), 338-342.

Burgess, P. W., Dumontheil, I., & Gilbert, S. J. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences*, *11*(7), 290-298.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage*, *14*(5), 1136-1149.

Dale, A.M., Fischl, B., Sereno, M.I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage, 9*, 179-194.

Davis, T., Goldwater, M., & Giron, J. (2017). From concrete examples to abstract relations: the

rostrolateral prefrontal cortex integrates novel examples into relational categories. *Cerebral Cortex, 27*(4), 2652-2670.

Davis, T., Love, B. C., & Maddox, W. T. (2009). Two pathways to stimulus encoding in category learning?. *Memory & cognition, 37*(4), 394-413.

Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex, 22*(2), 260-273.

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 821.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.

DeGutis, J., & D'Esposito, M. (2007). Distinct mechanisms in visual category learning. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(3), 251-259.

Deichmann, R., Gottfried, J. A., Hutton, C., & Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage*, *19*(2), 430-441.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... & Albert, M. S. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968-980.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23*(10), 475-483.

Ell, S. W., Marchant, N. L., & Ivry, R. B. (2006). Focal putamen lesions impair learning in rule-based,

but not information-integration categorization tasks. *Neuropsychologia*, *44*(10), 1737-1751.

Gilbert, S. J., Spengler, S., Simons, J. S., Steele, J. D., Lawrie, S. M., Frith, C. D., & Burgess, P. W. (2006). Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. *Journal of cognitive neuroscience, 18*(6), 932-948.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., ... & Rapoport, J. L. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National academy of Sciences of the United States of America*, *101*(21), 8174-8179.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*(3), 316-322.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual. Review of Neuroscience, 31*, 359-387.

Green, A. E., Fugelsang, J. A., Kraemer, D. J., Shamosh, N. A., & Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Research*, *1096*(1), 125-137.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63-72.

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, *49*(5), 757-763.

Heaton, R. K., Chelune, G. J., Curtiss, G., Kay, G. G., & Talley, J. L. (1993). *Wisconsin card sorting*

test. Psychological Assessment Resources.

Filoteo, J. V., Maddox, W. T., Simmons, A. N., Ing, A. D., Cagigas, X. E., Matthews, S., & Paulus, M. P. (2005). Cortical and subcortical brain regions involved in rule-based category learning. *Neuroreport*, *16*(2), 111-115.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*(18), 6117-6125.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(10), 2811-2822.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825-841.

Jenkinson, M., Beckmann, C. F., Behrens, T.E., Woolrich, M.W., Smith, S.M (2012). FSL. *NeuroImage, 62*, 782-790.

Konishi, S., Nakajima, K., Uchida, I., Kameyama, M., Nakahara, K., Sekihara, K., & Miyashita, Y. (1998). Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nature Neuroscience*, *1*(1), 80-84.

Konishi, S., Hayashi, T., Uchida, I., Kikyo, H., Takahashi, E., & Miyashita, Y. (2002). Hemispheric asymmetry in human lateral prefrontal cortex during cognitive set shifting. *Proceedings of the National Academy of Sciences*, *99*(11), 7803-7808.

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cerebral cortex, 12*(5), 477-485.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning.

Psychological Review, *99*(1), 22.

Liu, Z., Braunlich, K., Wehe, H. S., & Seger, C. A. (2015). Neural networks supporting switching, hypothesis testing, and rule application. *Neuropsychologia*, *77*, 19-34.

Ma, Q., Huang, B., Wang, J., Seger, C., Yang, W., Li, C., ... & Huang, R. (2016). Altered modular organization of intrinsic brain functional networks in patients with Parkinson's disease. *Brain imaging and behavior*, 1-14.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *The Journal of Neuroscience*, *21*(19), 7733-7741.

Nee, D. E., Jahn, A., & Brown, J. W. (2014). Prefrontal cortex organization: dissociating effects of temporal abstraction, relational abstraction, and integration with fMRI. *Cerebral Cortex*, *24*(9), 2377-2387.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145-8157.

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., ... & Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex, 17*(1), 37-43.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Peirce, J.W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods,*

162(1-2), 8-13.

Peirce, J.W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*:10. doi: 10.3389/neuro.11.010.2008

Poldrack, R. A., Clark, J., Pare-Blagoev, E. J., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546-550.

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience & Biobehavioral Reviews*, *32*(2), 197-205.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, *5*(3), 184-194.

Robinson, A. L., Heaton, R. K., Lehman, R. A., & Stilson, D. W. (1980). The utility of the Wisconsin Card Sorting Test in detecting and localizing frontal lobe lesions. Journal of consulting and clinical psychology, 48(5), 605.

Rorden, C., Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural Neurology, 12*, 191-200.

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J., & Verfaellie, M. (2009). Prefrontal contributions to rule-based and information-integration category learning. *Neuropsychologia*, *47*(13), 2995-3006.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience & Biobehavioral Reviews*, *32*(2), 265-278.

Seger, C. A., & Cincotta, C. M. (2006). Dynamics of frontal, striatal, and hippocampal systems during

rule learning. *Cerebral cortex*, *16*(11), 1546-1555.

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203.

Semendeferi, K., Teffer, K., Buxhoeveden, D. P., Park, M. S., Bludau, S., Amunts, K., ... & Buckwalter, J. (2011). Spatial organization of neurons in the frontal pole sets humans apart from great apes. *Cerebral Cortex*, *21*(7), 1485-1497.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., & Petersen, S. E. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high‐motion data points. *Human Brain Mapping*, *35*(5), 1981-1996.

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 249-264.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143-155.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83-98.

Specht, K., Lie, C. H., Shah, N., & Fink, G. R. (2009). Disentangling the prefrontal network for rule selection by means of a non‐verbal variant of the Wisconsin Card Sorting Test. *Human Brain Mapping*, *30*(5), 1734-1743.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W., & Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed

cognition. *Neuroimage*, *53*(1), 303-317.

Strange, B. A., Henson, R. N. A., Friston, K. J., & Dolan, R. J. (2001). Anterior prefrontal cortex mediates rule learning in humans. *Cerebral Cortex*, *11*(11), 1040-1046.

Therneau, T. (2015). coxme: Mixed Effects Cox Models. R package version 2.2-5. https://CRAN.R-project.org/package=coxme

Vendetti, M. S., & Bunge, S. A. (2014). Evolutionary and developmental changes in the lateral frontoparietal network: a little goes a long way for higher-level cognition. *Neuron, 84*(5), 906-917.

Wendelken, C., Chung, D., & Bunge, S. A. (2012). Rostrolateral prefrontal cortex: Domain‐general or domain‐sensitive?. *Human brain mapping, 33*(8), 1952-1963.

Wendelken, C., & Bunge, S. A. (2010). Transitive inference: distinct contributions of rostrolateral prefrontal cortex and the hippocampus. *Journal of cognitive neuroscience, 22*(5), 837-847.

Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in human neuroscience*, *5*.

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

Woolrich, M. (2008). Robust group analysis using outlier inference. *Neuroimage*, *41*(2), 286-301.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage*, *21*(4), 1732-1747.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., ... & Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, *45*(1), S173-S186.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, *14*(6), 1370-1386.

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.