A comparison of the neural correlates that underlie rule-based and information-integration category learning.

Kathryn L. Carpenter[1], Andy J. Wills[2], Abdelmalek Benattayallah[3] and Fraser Milton[1]

1. School of Psychology, University of Exeter, UK
2. School of Psychology, Plymouth University, UK
3. Exeter Medical School, University of Exeter, UK

Address for Correspondence: Kathryn L. Carpenter, Psychology, University of Exeter, Washington Singer, Exeter, EX4 4QG, United Kingdom. Tel: +44 (0)1392 724626. Email: klc206@exeter.ac.uk

Short Title: COVIS and category learning.

Key words: Magnetic Resonance Imaging; Learning; Learning, Verbal; Hippocampus; Parahippocampal Gyrus; Caudate Nucleus.

Manuscript Accepted at Human Brain Mapping

Abstract

The influential COmpetition between Verbal and Implicit Systems (COVIS) model proposes that category learning is driven by two competing neural systems – an explicit, verbal, system, and a procedural-based, implicit, system. In the current fMRI study, participants learned either a conjunctive, rule-based, category structure that is believed to engage the explicit system, or an information-integration category structure that is thought to preferentially recruit the implicit system. The rule-based and information-integration category structures were matched for participant error rate, the number of relevant stimulus dimensions and category separation. Under these conditions, considerable overlap in brain activation, including the prefrontal cortex, basal ganglia, and the hippocampus, was found between the rule-based and information-integration category structures. Contrary to the predictions of COVIS, the medial temporal lobes and in particular the hippocampus, key regions for explicit memory, were found to be more active in the information-integration condition than in the rule-based condition. No regions were more activated in rule-based than information-integration category learning. The implications of these results for theories of category learning are discussed.

Category learning is an essential cognitive process necessary for daily functioning. Without the ability to categorize an object as a threat, for instance, our survival chances would be severely impeded. But how do we learn novel categories? For example, how does a student driver learn to categorize the symbols on the road? Poldrack and Foerde (2008) regard the concept of multiple memory systems as one of the most important contributions to neuroscience in the past quarter century, and one increasingly prominent line of research within this broader field is the idea that there are multiple systems of category learning (Ashby & Maddox, 2011). Perhaps the most influential multiple systems account is the dual-process, neurobiologically inspired, COmpetition between Verbal and Implicit Systems (COVIS) model of category learning (Ashby, Alfonse-Reese, Turken & Waldron, 1998) which is the focus of the current study.

COVIS hypothesizes that there are two neurally and functionally dissociable category learning systems (Ashby et al. 1998). The explicit system requires considerable use of working memory and executive functioning to test the effectiveness of rules that are generated. This learning system, consequently, works best at learning rule-based (RB) category structures where the decision boundary separating the categories can be easily verbalized. The most common examples in the literature are unidimensional rules such as "short lines belong in category A; long lines belong in category B" (see Figure 1a), or conjunctive rules such as "short, upright, lines belong in category A; anything else belongs in category B" (see Figure 1b). On the other hand, in the implicit, procedural-based, system, learning occurs by combining information from two or more unrelated stimulus dimensions predecisionally through reliance upon immediate feedback to create stimulus-response associations (Ashby et al. 1998). The implicit system is usually tested using information-integration (II) categories (see Figure 1c) where the optimal decision boundary is typically considered difficult or impossible to verbalize.

One notable aspect of COVIS that distinguishes it from other multiple system accounts of category learning (e.g., ATRIUM, Erickson & Kruschke, 1998; RULEX, Nosofsky, Palmeri, & McKinley, 1994) is the detailed neurobiological predictions that it makes regarding the brain regions that underlie the different learning systems. In the explicit system, rule generation and hypothesis testing requires working memory and executive functioning which takes place predominately in the prefrontal cortex (Ashby & Valentin, 2005). The particular rule to use is selected via the anterior cingulate (Maddox & Ashby, 2004), while the head of the caudate nucleus is responsible for mediating the switch to a different rule. Successful rules are stored in the medial temporal lobes (MTL) for future use (Ashby & Valentin, 2005). The MTL is also hypothesized to store representations of the decision boundaries used to separate the stimuli into categories (Nomura & Reber, 2008).

In contrast, the implicit system procedurally acquires the stimulus-response associations necessary for learning II categories (Ashby et al., 1998). The body and tail of the caudate nucleus receive representations of the visual stimulus perceived (Ashby & Valentin, 2005) and these cells project to the supplementary motor area via the globus pallidus and the thalamus (Maddox & Ashby, 2004). When feedback indicates a correct response has been made, the substantia nigra releases dopamine which strengthens the association of the stimulus to the correct response (Ashby & Valentin, 2005). The putamen has also recently been proposed by Waldschmidt and Ashby (2011) to play a key role in the implicit system, as it is assumed to provide information to the motor regions (but see Ell, Marchant, & Ivry, 2011, who found that focal putamen lesions impaired RB but not II learning).

These neurobiological underpinnings of COVIS have motivated a large number of predictions about how RB and II learning will be differentially affected by behavioral manipulations (for reviews see Ashby & Maddox, 2011; Maddox & Ashby, 2004). The numerous behavioral dissociations arising out of this work have contributed a great deal to

the influence of COVIS. For example, RB learning is impaired by the imposition of a concurrent working memory load while II learning is not, supporting the idea that working memory is more critical for RB than II learning (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). Similarly, II learning is disrupted by changing the appropriate response buttons while RB learning is not, in line with the prediction that II learning relies upon stimulus-response procedural associations (Ashby, Ell & Waldron, 2003). Furthermore, delaying feedback for a few seconds (Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005), deferring feedback to the end of a block of 6 trials (Smith et al., 2014), providing the category label prior to making the response (Ashby, Maddox, & Bohil, 2002), and using stimuli which contain both auditory and visual information (Maddox, Ing, & Lauritzen, 2006) have all been claimed to impair II but not RB learning. Similarly, some studies have suggested that increasing the number of categories (Maddox, Filoteo, Hejl, & Ing, 2004) or reducing the time available to process the feedback (Maddox, Ashby, Ing, & Pickering, 2004) disrupts RB but not II learning.

However, in recent years there have been a growing number of studies that cast doubt on COVIS's interpretation of these behavioral dissociations and posit that the results can be explained by a single, explicit, system (e.g., Nosofsky & Kruschke, 2002; Newell, Dunn, & Kalish, 2010, 2011; Newell et al.,2013; Stanton & Nosofsky, 2007, 2013). One such example is a study conducted by Lewandowsky, Yang, Newell, and Kalish (2012) who reconsidered the finding noted above that RB learning relies on working memory to a greater extent than II learning. Specifically, Lewandowsky et al. directly measured the working memory capacity of participants using a battery of both verbal and spatial tasks and used structural equation modeling to reveal a strong relationship between working memory capacity and both RB *and* II learning, consistent with their proposal that both tasks require the use of working memory. The behavioral evidence as it currently stands, therefore, provides equivocal support for

COVIS. An alternative approach to resolving this dispute is to focus directly on the neurobiological predictions of COVIS where it has been argued that single-system accounts cannot explain the evidence that separable neural systems are engaged during different types of category learning (Worthy, Markman & Maddox, 2013).

This neurobiological evidence is currently surprisingly limited, however, as there has been a paucity of studies directly comparing the brain systems involved in RB and II category learning. Perhaps the most prominent study to examine this, though, was by Nomura et al. (2007). Participants completed either an RB or an II category learning task inside an MRI scanner. The RB category structure was an easy to verbalize unidimensional rule (e.g., Figure 1a), while the II structure was based on that shown in Figure 1c. Nomura et al. considered their results to be in line with COVIS - dissociable neural activation was found, with the MTL more activated in RB compared with II learning, and the caudate body more activated in II than RB learning. Further evidence of separable systems was found in a reanalysis of Nomura et al.'s data which modeled participants' decision strategies (Nomura & Reber, 2008). Participants using RB learning strategies showed greater right PFC activity than those using II strategies, and those utilizing II strategies had greater right occipital activation.

More recently, Soto, Waldschmidt, Helie and Ashby (2013; see also Helie, Waldschmidt, & Ashby, 2010; Waldschmidt & Ashby, 2011) directly contrasted RB and II learning in a multi-voxel pattern analysis. While the study had multiple training sessions, the first scanning session (Training session 1 for the RB task and Training session 2, following 600 training trials in Session 1, for the II task), prior to the development of automaticity, is most pertinent for the current issue. While there was common activation between RB and II learning (for example in the globus pallidus and the extrastriate visual cortex), there were some differences in activation. For instance, consistent with Nomura et al. (2007), the head of the caudate was activated more in RB learning, while activation in the caudate body/tail also

differed between the RB and II tasks. However, it is difficult to know whether these neural differences were due to the engagement of separate systems in RB and II learning or whether they are due to participants in the II condition having already received 600 training trials previous to the scanning session while participants in the RB condition had no prior training (brain activation alters over a relatively limited number of trials, e.g., Koenig et al., 2005; Milton & Pothos, 2011). While this issue was not the sole focus of Soto et al.'s study it does, nevertheless, compromise any direct comparisons in brain activation between II and RB learning prior to automaticity developing.

Milton and Pothos (2011) found a different pattern of results to Nomura et al. (2007), observing extensive overlap in activation between a Unidimensional RB structure and a Complex category structure assumed to have many of the properties of II categories (e.g., optimal decision bounds that were difficult to verbalize). In contrast to Nomura et al., neural differences between the II and RB conditions were minimal and restricted to greater activation in a small region of the left superior frontal lobe in the Complex condition relative to the RB condition. While intriguing, one should not draw too strong an inference about these findings with regard to COVIS due to the differences in the stimuli that Milton and Pothos used compared to those traditionally administered in COVIS research. For instance, there were only 18 unique stimuli, with dimensions that were commensurable (rectangle height and ellipse width) and a decision bound that was arguably easier to verbalize than the II structures typically employed (e.g., Figure 1c). Nevertheless, these findings indicate that further direct comparison of the neural correlates of RB and II category learning is needed.

The aim of the present study, therefore, is to re-examine Nomura et al.'s (2007) conclusion that there is a differential pattern of brain activation for RB and II categories in line with the predictions of COVIS. The critical difference between the RB and II category structures is often assumed to be that the RB structure is easily verbalizable but the II

structure is not. While the RB and II category structures used by Nomura et al. (see Figures 1a and 1c) differ convincingly in this factor, there are also non-essential differences between them that may potentially be driving the differences in activation. For instance, the RB structure has only one relevant dimension while the II structure has two relevant dimensions which means that selective attention is required for the RB condition but not for the II condition (Nosofsky & Kruschke, 2002). This is a concern that has been acknowledged by some COVIS theorists (e.g., Nomura & Reber, 2008; Zeithamova & Maddox, 2006; Xie, Maddox, McGeary, & Chandrasekaran, 2015). On a different note, multi-dimensional categorizations are typically more complex and require greater cognitive resources than one-dimensional categorizations (e.g., Milton, Longmore, & Wills, 2008; Wills, Inkster, & Milton, 2015). This could potentially be driving the more pronounced caudate body activation in the II than the RB condition, particularly given that the involvement of the basal ganglia is thought to be greater for more complex structures (e.g., Ell, Weinstein, & Ivry, 2010; Filoteo, Maddox, Salmon & Song, 2005). As the II structure is often more difficult to learn than the RB structure (e.g., Ashby, Maddox, & Bohil, 2002, Maddox, Ashby, & Bohil, 2003), Nomura et al. reduced the category separation (i.e., the mean distance between category items as plotted in stimulus space divided by the within-category variance along the direction of the comparison) in the RB condition relative to the II condition to minimize any performance differences between conditions (see also Lewandowsky et al. 2012 for a discussion of this issue). While this successfully matched learning rates, it effectively replaces one confound with another because the optimal decision bound is more difficult to perceptually discriminate in the RB than the II condition (Stanton & Nosofsky, 2007). This confound is potentially critical given that COVIS assumes that the MTL is responsible for storing the precise placement of the decision bound (Nomura & Reber, 2008). The greater

activation in the MTL for the RB condition compared to the II condition could, therefore, be due to this difference in category separation.

In order to draw strong comparisons about brain activation in RB and II category learning, it is therefore necessary to control for these non-essential differences between the category structures. This has been achieved in previous COVIS related research (e.g., Filoteo et al., 2010; Zeithamova & Maddox, 2006) - but in no previous imaging study - by comparing the II category structure to a conjunctive, rule-based category structure (see Figure 1b). The II and the conjunctive category structures both possess two relevant dimensions, have a similar error rate (Filoteo et al., 2010) and are closely matched for category separation.

A study by Edmunds, Milton and Wills (2015) underscores the importance of controlling for these extraneous variables when comparing RB and II category learning. Edmunds et al. re-examined Ashby et al.'s (2002) finding that trial-by-trial feedback training leads to better categorization performance than observational training for an II structure but not for a unidimensional RB structure. Edmunds et al. argued that Ashby et al.'s dissociation could have been driven by one of the non-essential differences between these category structures highlighted above. In particular, Edmunds et al. posited that the increased difficulty of learning a multi-dimensional (II) classification compared to a unidimensional classification could have been causing the effect - feedback training may be of greater benefit than observational training more generally but this advantage increases as the problem difficulty rises. To investigate this, Edmunds et al. compared learning of conjunctive and II category structures under both observational and feedback training. Edmunds et al. confirmed that participants are better able to verbalize the conjunctive, RB, structure than the II structure. However, the dissociation predicted by COVIS failed to emerge. Instead, feedback learning was superior to observational learning for both category structures.

The importance of controlling important extraneous variables has also been highlighted in a recent fMRI study conducted by Nosofsky, Little and James (2012) who reconsidered the classic finding of Reber, Stark & Squire (1998) that old-new recognition of dot patterns evokes a different pattern of brain activation to categorization of dot patterns. Nosofsky et al. used the same stimuli across conditions (whereas in Reber et al. the stimuli differed) and more closely equated the task goals of recognition and categorization (normally recognition requires an exact match with the studied item while stimuli can be endorsed as a category member if they are merely similar to previous exemplars) by asking them to adopt a lax criterion for the recognition judgment - participants were told it was important not to miss any old items. Under these conditions, there was little evidence for dissociable systems and the results from both tasks could be accommodated by a single exemplar-based process.

Another notable aspect of Nomura et al.'s (2007) study is their use of incorrect trials as the baseline comparison to correct responses. While this is a convenient baseline to use and has been employed in other categorization research (e.g., Milton & Pothos, 2011) it may not be the most effective due to difficulties in interpreting what is driving the incorrect response. First, participants may have been using the correct general strategy but had not identified the relevant dimension/precise category structure; for example, participants used a rule-based strategy but categorized by orientation rather than line length. Second, participants might have used the appropriate dimension but placed the decision bound in the incorrect place. Third, and less commonly, participants may have classified correctly but pressed the wrong button. Fourth, participants may have been guessing or not fully engaged on the trial and fifth, participants could have used a completely different strategy to what was appropriate. It is likely that the errors are a combination of these (and potentially other) mistakes but it is not possible at the individual trial level to determine the source of the error. The first three of these error types appear particularly problematic as they would result in

similar activation to correct trials meaning that this is unlikely to be a sensitive baseline. Furthermore, comparing correct and incorrect trials is likely to be confounded with degree of learning as there will be more incorrect trials early in training than later in training. This is particularly an issue when wishing to make inferences across the whole of training as is typically the case. While we present the key analyses with this "incorrect" baseline to aid comparison of our results with Nomura et al.'s, an "odd-or-even" task will be our principal baseline. This type of control is increasingly being used in imaging studies of categorization (e.g., Davis, Love & Preston, 2012a, 2012b; Davis, Xue, Love, Preston & Poldrack, 2014) and while it may superficially seem similar to a RB task (albeit one that is highly automated and engages limited neural resources, Stark & Squire, 2001) its main advantage is that it is well-established that it does not recruit the MTL or indeed the frontal lobes (Stark & Squire, 2001), the pivotal regions of COVIS's rule-based system. Equally, activation in the striatum, the key site of COVIS's implicit system, is also readily identified with an odd-or-even baseline task (Zink et al., 2006). This baseline should, therefore, provide a clear measure of the regions engaged in both RB and II categorization without the involvement of key regions being obscured by their activation in the baseline task as well.

According to how COVIS is often conceptualized (e.g., Nomura et al., 2007), one might predict greater activation in the caudate head, the anterior cingulate, prefrontal cortex, and the MTL (and in particular the hippocampus) for learning a conjunctive RB structure compared to learning an II structure (Ashby & Valentin, 2005). In contrast, greater activation should be found in the body/tail of the caudate, the putamen and the substantia nigra for the II condition compared to the RB condition (Ashby & Valentin, 2005). Conversely, if Nomura et al.'s (2007) results were driven by one of the non-essential differences between the RB and II structures outlined above then, when these variables have been better controlled, one might expect that these neural differences would disappear leaving an extensive overlap of

activation. Further, given that the II structure is harder to verbalize than the RB structure and yet categorization accuracy is the same (Filoteo et al., 2010), greater activation might be expected in the prefrontal cortex for the II compared to the RB condition to reflect the greater processing demands of finding and applying a less easy to verbalize rule. Alternatively, or perhaps additionally, there may be greater activation in the MTL in the II condition than the RB condition if the lower levels of verbalizability lead to an increase in memory demands to store exceptions in decision space to the rule that is utilized (Davis et al., 2012a).

Method

Participants and Design

45 right-handed University of Exeter students (26 female, 19 male) with normal or corrected vision completed the experiment for £5 remuneration. Participants were randomly allocated to one of two between-subject conditions (RB and II). One participant from the RB condition was excluded for failing to reach 50% (chance) accuracy in the final run (although the inclusion of this participant does not alter any of the conclusions of this study), leaving 22 participants in each condition. Participants gave informed consent according to procedures approved by the University of Exeter's School of Psychology Ethics Committee.

Stimuli

The stimuli (see Figure 2) were a subset of the two-dimensional II stimuli and conjunctive stimuli (where short, upright lines belong in category A, and the rest in category B) employed by Filoteo et al. (2010). In the original data set there were 600 stimuli in both conditions; in the present imaging study, 320 of these stimuli were randomly selected (160 stimuli in each category) for each category structure. This number of stimuli was the same as used by Nomura et al. (2007). Each stimulus was a black line varying on two dimensions: length and orientation. As in Filoteo et al. (2010), there was 5% overlap between the categories so that the maximum accuracy attainable was 95%.

fMRI imaging

A 1.5-T Phillips Gyroscan magnet, equipped with a Sense coil, was used to collect images from each participant in one scanning session. A T2*-weighted echo planar sequence (TR = 3000ms, TE = 45ms, flip angle = 90°, 36 transverse slices, 3.5 x 2.5 x 2.5mm) was used. Upon entering the scanner, the participant's head was secured in place with foam pillows inside the coil to prevent excessive head movement. Participants completed four runs, each containing 205 scans. Five "dummy scans" were completed before every run prior to presentation of the first trial. After the functional scans, standard volumetric anatomical MRI was completed using a 3-D T1-weighted pulse sequence (TR = 25ms, TE = 4.1ms, flip angle = 30°, 160 axial slices, 1.6 x 0.9 x 0.9mm).

Procedure

In each scanning run, participants performed two interleaved tasks - the category learning task and an "odd-or-even" baseline task. Each run began with 15 odd-or-even trials, followed by two blocks of 40 categorization trials. Each run then concluded with another block of 15 odd-or-even trials. After each block there was a blank screen of 8000ms during which time participants were asked to rest. In total, there were 320 category learning trials, presented in a random order, and 120 odd-or-even trials. The stimuli were presented on a back-projection screen positioned at the foot end of the MRI scanner and viewed via a mirror mounted on a head coil. Responses were measured using a fiber-optic button box held in the participants' left and right hands. E-Prime (Psychological Software Tools, 2002) was used for the presentation and timing of stimuli and collection of response data.

In the *category learning task*, participants were informed that they had to learn into which of two categories a series of stimuli belonged. The trial-by-trial procedure for the RB and II conditions was identical. Each trial began with a blank screen lasting a variable

interval between 500ms and 4000ms, followed immediately by a black fixation cross presented in the center of the screen for 250ms. A stimulus then appeared in the middle of the screen for 2000ms during which time participants were required to respond by pressing the far right button on the button box with their right hand if they thought the item belonged to category A or the far left button with their left hand if they thought the item was a member of category B. Feedback ("Correct" or "Incorrect") was then displayed for 500ms. If participants did not respond in time the message "Time out!!!" appeared on the screen for 500ms instead. The next trial then immediately began.

The odd-or-even task was closely modeled on that used by Stark and Squire (2001; see also Davis et al., 2012a) and had a similar trial-by-trial structure to the category learning task. Each trial began with a blank screen lasting between 500-4000ms, followed by a black fixation cross for 250ms. A randomly generated number from one to nine then appeared in the middle of the screen for 2000ms during which time participants had to press the left-most button if the number was even or the right-most button if it was odd. Following this, feedback ("Correct" or "Incorrect") was presented for 500ms or if participants did not respond in time a message saying "Time out!!" appeared during this interval.

fMRI Data Analysis

Data analysis was performed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm). Functional images were corrected for acquisition order, realigned to the mean image, and resliced to correct for motion artifacts. The realigned images were coregistered with the structural T1 volume and the structural volumes were spatially normalized. The spatial transformation was applied to the T2* volumes which were spatially smoothed using a

Gaussian Kernel of 8mm full-width half maximum. Data were high-pass filtered (128s) to account for low frequency drifts.

Random effects whole-brain analyses were completed using the general linear model with a combined statistical threshold of $p<.001$ (uncorrected) and a voxel threshold of 27 contiguous voxels, which together produce an overall corrected threshold of $p<.05$, according to AlphaSim, as implemented in the REST toolbox (Version 1.8; Song et al., 2011). Correct trials, incorrect trials, and timeouts were all included as separate regressors in the model. A canonical hemodynamic response function (HRF) together with temporal and dispersion derivatives was used to model the blood oxygen level-dependent response and the six head movement parameters were included as covariates. Our analyses focused on comparing correct categorization trials (for the RB and II groups separately) to the odd-or-even baseline task (although for the principal analyses we also compare correct categorization trials to incorrect trials). In addition, to measure common activation between the RB - baseline contrast and the II - baseline contrast, a conjunction analysis was performed. The contrasts were combined using a logical 'and' function through the minimum statistic to the conjunction null hypothesis (MS/CN; Nichols et al., 2005) technique implemented in SPM8. Both these contrasts were again conducted with a combined threshold of $p<.001$ (uncorrected) and a cluster threshold of 27; note that this approach is highly conservative because it reveals only those regions significantly activated for both the RB ($p<.05$, corrected) *and* the II ($p<.05$, corrected) conditions. Normalized MNI space coordinates were transformed to Talairach space (http://imaging.mrccbu.cam.ac.uk/imaging/MniTalairach) to establish activation sites as per the atlas of Talairach and Tournoux (1988).

Results

Behavioral Analysis

The mean categorization accuracy across all runs for both the RB and II conditions is displayed in Figure 3. A 4 x 2 mixed-design analysis of variance (ANOVA) was conducted; the within-subjects factor was Run (4 levels) and the between-subjects factor was Categorization task (RB/II). There was a highly significant effect of Run, $F(3,126)=12.47$, $p<.001$, $\eta^2_p = .229$, indicating that performance improved with practice. There was, however, no significant difference between the II and RB conditions in accuracy $F(1,42)=.14$, $p=.708$, $\eta^2_p = .003$, BF = 1.04, and no significant interaction between Run and Categorization task (II or RB), $F(3,126)=1.71$, $p=.169$, $\eta^2_p .= .039$, BF = 2.00[1].

Imaging Analysis - 'Odd or Even?' Baseline Measure

All blocks analysis

Whole-brain activation across all runs of the category learning task was first analysed for participants in the RB and II conditions separately. Correct RB categorizations led to an extensive pattern of activation (Figure 4a) including diverse areas of the frontal cortex (including BA's 6, 8, 10, 45, 46, 47), the anterior cingulate, posterior cingulate, the MTL, the bilateral caudate head/body, the putamen, the bilateral inferior and superior parietal lobes, the right superior temporal gyrus, bilateral inferior temporal gyrus and the bilateral occipital lobes. II category learning also activated these same brain regions (Figure 4b).

We also examined whether there were any changes in activation across time for both RB and II learning. To assess this, we directly compared activation in the first half of the

experiment (runs 1 and 2) to in the second half (runs 3 and 4) for the RB and II conditions separately. No brain regions were more activated in the first half of training compared with the second half of training in either condition. No brain areas were activated more in the second half of training than the first half in the RB condition either. However, in the II condition several regions including the right parahippocampal gyrus (BA 30; see Table I) were activated more in runs 3 and 4 than in runs 1 and 2.

The striking overlap in activation between the tasks was confirmed in a conjunction analysis, looking at common activation across the correct RB - odd-or-even contrast and the correct II - odd-or-even contrast (both with thresholds of $p<.001$ and 27 contiguous voxels; Figure 4c). Areas activated included key regions of both COVIS's explicit and implicit systems. Regions linked to the explicit system that were recruited were the MTL, the bilateral caudate head, diverse bilateral areas of the prefrontal cortex (including BA's 6, 8, 10, 46, and 47) and the bilateral anterior cingulate (right BA 25, left BA 33). Areas implicated in the implicit system that were engaged included the bilateral caudate body and the bilateral putamen. When contrasting incorrect trials to the odd-or-even task, a similar, if somewhat less extensive, pattern of activation was found including the right caudate body, right putamen and bilateral caudate head (Supplementary Figure 1).

Next, we directly contrasted activation in the RB and II conditions to examine whether there was evidence for the neural dissociations observed by Nomura et al. (2007). No regions were more active in the RB condition than the II condition (calculated by subtracting correct RB trials - the odd-or-even trials from correct II trials - the odd-or-even trials). However, diverse regions were more active in the II condition than the RB condition (calculated by subtracting correct II trials - the odd-or-even trials from correct RB trials - the odd-or-even trials; see Table II, Figure 5a). Critically, this included extensive activation in the left MTL (hippocampus/ posterior parahippocampal gyrus; 131 voxels; see Figure 5b for

areas of the MTL engaged, with non-MTL regions masked). The results of these analyses are contrary to the predictions of COVIS, where the MTL is thought to be more critical for RB rather than II learning (e.g., Ashby & Valentin, 2005; Nomura et al., 2007).

However, in spite of the generally greater activation in the II condition than the RB condition, no regions associated with COVIS's implicit system were identified in this analysis. Of course, it is possible that, even though we had almost double the number of participants that Nomura et al. (2007) used (they had 13 in their II condition and 12 in their RB condition), this activation might have been present but below our a priori statistical thresholds. To provide greater sensitivity we, therefore, conducted a region of interest (ROI) analysis using the WFU Pickatlas (Maldjian, Laurienti, Burdette, & Kraft, 2003) comprising the caudate body, the putamen, and the substantia nigra with the more liberal thresholds of $p<.005$ and 10 contiguous voxels (the same thresholds we have used in previous ROI analyses we have conducted, c.f., Milton et al., 2011; Milton, Butler, Benattayallah, & Zeman, 2012). We again found no evidence for greater activation in the II condition than the RB condition in these regions. To further confirm this conclusion we examined the relative percent signal change of correct RB and II responding in the caudate body based on the peak right (x = 17, y = -11, z = 28) and left (x = -20, y = -14, z = 29) caudate body activations reported by Nomura et al. (2007). These percent signal change values were obtained using the Anatomy toolbox (Eickhoff et al., 2007, Version 2.2). Using independent samples t-tests, we found no difference between conditions for either the right caudate body, $t(42)=1.05$, $p=.300$, $d= .32$, $BF=.84$, or for the left caudate body, $t(42)=1.00$, $p=.323$, $d= .3$, $BF=.91$[2].

Analysis of runs 3 and 4 only

COVIS can potentially explain this pattern of findings by assuming that for the II condition as well as the RB condition the verbal system dominates initially and participants in the II group only switch to the implicit system once there has been sufficient time for the RB system to be proven ineffective (e.g., Filoteo et al., 2010). Including the initial trials in the analysis could therefore be obscuring the neural differences that emerge later in learning. To investigate this possibility, we analyzed runs 3 and 4 alone which, according to the results of previous studies (e.g., Filoteo et al., 2010), should provide a sufficient number of trials for participants to switch to the implicit system in the II condition.

A conjunction analysis, using the same thresholds as before, again revealed extensive activation overlap between the RB and II conditions. This included the bilateral putamen, the bilateral caudate body as well as the bilateral caudate head, the prefrontal cortex and the right MTL (Supplementary Figure 2). No regions were again more activated in the RB condition than the II condition. However, as before, a number of regions were more activated in the II condition than the RB condition (Supplementary Table I; Figure 6a); most prominent amongst these was activation in the bilateral hippocampus/posterior parahippocampal gyrus (left: 207 voxels; right: 44 voxels, Figure 6b). However, as before, in spite of this generally elevated activation in the II condition compared to the RB condition, there was no evidence for activation of regions linked to COVIS's implicit system. We again conducted a follow-up ROI analysis comprising the caudate body, the putamen and the substantia nigra with a threshold of $p < .005$ and a voxel threshold of 10 but no regions were activated in this analysis.

Model based analysis

The predictions made by COVIS are, of course, dependent on the assumption that more participants in the RB condition are using the explicit system than are participants in the

II condition and that a greater number of participants in the II condition are using the implicit system than are participants in the RB condition. If participants in the II condition persist with the verbal system throughout (or alternatively if participants in the RB condition as well as the II condition use the implicit system) then this might explain why our results appear inconsistent with the predictions of COVIS. It is harder, though, from a COVIS perspective to explain why participants in the II condition engaged the MTL, a critical region of the explicit system (Ashby & Valentin, 2005; Nomura & Reber, 2008), more than participants in the RB condition unless one assumes that the II category structure was more effective than the RB structure at engaging the explicit system. While this may seem unlikely, it can be tested using model-based analysis based on General Recognition Theory (GRT; Ashby & Gott, 1988) as is commonly carried out in COVIS related studies (e.g., Ashby et al., 2002; Filoteo et al., 2010; Nomura & Reber, 2008).

For each participant, the GRT analysis determines the decision boundary (from a set of pre-defined alternatives) that provides the best account of that participant's responses. Each participant is then assigned a strategy type (e.g. 'conjunctive') on the basis of the best-fitting model.

The *unidimensional* models assume that the participant determines a criterion along either the orientation or length dimension. As an example, for length, this corresponds to a rule such as: 'Assign to Category A if the stimulus is long, or Category B if short'. The unidimensional models have two parameters: the value of the criterion and the variance of internal (criterial and perceptual) noise.

The *conjunctive* model assumes that the participants make two judgments, one for each stimulus dimension, and then combine these to make a judgment about category membership. The conjunctive rule in the current analysis was: 'Assign the stimulus to

Category A if it is short and upright, otherwise assign to Category B'. The conjunctive model has three parameters: the two criterion values and internal noise.

The *General Linear Classifier (GLC)* model assumes that the decision boundary can be described by a straight line that can vary in gradient and intercept. The unidimensional models are therefore special cases of the GLC model. The GLC model has three parameters: the intercept and slope of the decision bound, plus internal noise.

The *random* model assumes that participants are responding randomly; it has no parameters.

For each participant, the best fit of each of these models was calculated, and the best-fitting model selected using Akaike's information criterion (Akaike, 1974). The results from this analysis, which was performed using the grt package in the R environment (Matsuki, 2014), are reported in Table III. Within the COVIS framework, the unidimensional and conjunctive models are considered to represent explicit, rule-based strategies, while the GLC represents an implicit, information-integration strategy.

The results, displayed in Table III, are generally consistent with previous work indicating that more participants used a conjunctive strategy in the RB condition than in the II condition and that more participants in the II condition used a GLC strategy than in the RB condition. This is the pattern expected and obtained in previous COVIS studies (e.g., Ashby et al., 2002); therefore, the modeling analyses seem to rule out the possibility (at least within the COVIS framework) that our results were driven by participants not using the intended strategy for their condition. However, the GRT modeling results, as usual, indicate that not all participants are adopting the expected strategy, so we took those participants in the RB condition whose responses were best fit by a conjunctive strategy (12 participants) and compared their brain activation to those participants in the II condition whose responses were

best fit by the optimal GLC model (13 participants). Note, the selection of a subset of participants on the basis of their GRT modeling results has seldom previously been carried out in COVIS related studies, perhaps because there are limits to the accuracy of these modeling results (see Donkin et al., 2015; Edmunds et al., 2015, for a discussion) so this analysis should be taken with some caution. Nevertheless, given the nature of our results, these supplementary analyses appear valuable.

A conjunction analysis, using, as before, thresholds of $p<.001$ and 27 contiguous voxels, again revealed an extensive overlap of activation between the RB and II conditions in similar regions to those found in the whole-group analyses (Figure 7a). Regions activated included the left MTL, bilateral caudate head, as well as the bilateral caudate body and right putamen. As in the all-participant analyses, no areas were more active in the RB than in the II condition. No regions were activated more in the II than the RB condition either and, in particular, the prominent MTL activation in the all-participant analyses did not emerge. One potential reason for this is simply that the smaller number of participants in this model-based analysis reduced our ability to detect this activation. We therefore conducted a post-hoc ROI analysis of the MTL (using the WFU Pickatlas; Maldjian et al., 2003) with a threshold of $p<.05$ (uncorrected) and a cluster threshold of 79 (which combined produce a corrected threshold of $p<.05$ according to AlphaSim). This revealed activation in the same left hippocampus/ parahippocampal gyrus region (cluster size: 115; Figure 7b) as previously identified.

We again found no activation in regions linked to COVIS's implicit system in the II - RB analysis. We therefore conducted another post-hoc ROI analysis comprising the caudate body, substantia nigra and the putamen in the same manner as for the MTL ROI analysis with cluster thresholds of $p<.05$ (uncorrected) and 41 contiguous voxels (which corresponded to $p<.05$, corrected according to AlphaSim). This also did not produce any significant

activation. An additional ROI analysis with these regions using alternative thresholds of $p<.005$ and 10 voxels also yielded no activation. Finally, we repeated these modeling analyses with runs 3 and 4 alone. These produced the same pattern of results as the all-run analyses - there was considerable common activation (see Supplementary Table II) with no regions more activated in the RB compared with the II condition (RB – II). There was, though, as in the corresponding all-participants analysis, evidence of left MTL activation in a post-hoc ROI analysis of the II - RB contrast with thresholds of $p<.05$ and 79 contiguous voxels (cluster size: 157, peak voxel: x = -12, y = -41, z = 4), and no evidence for COVIS's implicit system in either the whole-brain or ROI analyses.

Correct – Incorrect trials

To complement the analyses just described we also ran the principal ones using incorrect trials as the baseline. For the all-blocks analysis, consistent with previous work (e.g., Cincotta & Seger, 2007; Filoteo et al., 2005), we found that the left caudate head was more active on correct trials than incorrect trials (with thresholds of $p<.001$ and 27 contiguous voxels) for both the RB (peak voxel: x = -16, y = 20, z = 5) and II (peak voxel: x = -8, y = 13, z = -6) groups. We again identified, using the same conjunction analysis approach as before, large overlap of activation between the II and RB conditions including bilateral putamen, left caudate body, right MTL and frontal lobe (including BA 8, 9, 10 and 11) (Table IV). We did not, however, detect any differences between II and RB learning. A similar pattern emerged when considering all blocks in the modeling analysis with activation overlap in the conjunction analysis, but no differences detected between II and RB learning.

Looking at blocks 3 and 4 alone, there was again common activation in the frontal, parietal and temporal lobes (Supplementary Table III) and no differences between RB and II

learning in whole-brain analyses. However, in a similar ROI MTL analysis to before (thresholds p=.05 and 79 contiguous voxels, corresponding to p<.05, corrected), we observed greater activation in the II condition than the RB condition in two right hippocampus/parahippocampal gyrus regions (cluster size 215, peak coordinate: x = 24, y = -7, z = -13; cluster size 180, peak coordinate: x = 34, y = -34, z = -12; Supplementary Figure 3), with the posterior cluster being in the same area as observed in the corresponding odd-or-even comparison. There was, though, again no evidence for activation in regions associated with COVIS's implicit system even when the analogous ROI analyses to those previously conducted were performed. This same pattern emerged when considering blocks 3 and 4 alone in the modeling analysis.

Discussion

Previous work has found that there is differential brain activity in the learning of RB and II categories with the MTL preferentially recruited for RB compared to II learning while the caudate body is engaged more for II than RB learning (Nomura et al., 2007). However, we found no evidence for this pattern of dissociable neural activation. In particular, our most noteworthy finding was that the hippocampus/posterior parahippocampal gyrus was significantly more activated in the II condition than the RB condition. In addition, there was a striking overlap of activation between RB and II category learning emphasizing the extensive common neural processes that are engaged in learning both category structures. Common activation included regions thought to be engaged both in the explicit system such as the prefrontal cortex (including BA's 8, 10, 46, and 47), the anterior cingulate, the caudate head, and the MTL, and regions implicated in the implicit system including the posterior caudate, the putamen, and the substantia nigra. This pattern persisted, and indeed became more pronounced, when the second half of training was analyzed alone. We also observed the same basic findings when including only those participants who had used the intended strategy as indicated by GRT modeling analyses.

The striking overlap in activation between RB and II learning is consistent with the idea that RB and II category learning require similar neural processes. Of course, some of this activation is likely to be related to processes not specific to the act of categorization itself but common functions shared by the tasks such as stimulus processing, response selection, feedback monitoring, uncertainty and attentional demands to name a few possibilities.

A somewhat related way of looking at this common activation is that the behavioral dissociations in past work (e.g., Ashby & Maddox, 2011, but see also Newell et al., 2011) may reflect true differences in the learning system engaged in RB and II categorization but

that the functions of these systems share similar neural pathways. For example, Duncan (2010) proposed that there is a multiple-demand brain network, comprising regions of the prefrontal and parietal cortex, that is responsible for integrating and coordinating the processing of task specific brain areas through dividing the task goal into sub-tasks, generating the rules to achieve each sub-goal and transferring information from one task to another. If such a system organized the specific separable processes needed for learning in the RB or II condition then neural activation overlap (such as in the frontal lobes) between these tasks would be apparent as seen in the present study.

Nevertheless, what is particularly striking about our results and a challenge to COVIS, as it is currently formalized, is that we found extensive regions of the hippocampus/posterior parahippocampal gyrus were activated *more* in the II condition than the RB condition when COVIS appears to make the reverse prediction that there should be *less* activation in the II condition than the RB condition in this region.

One important question, therefore, is why we observed a markedly different pattern of results from Nomura et al. (2007)? It appears unlikely that this is due primarily to our choice of the odd-or-even task as our primary baseline measure because the same basic pattern of results was observed when we used incorrect trials as the baseline (albeit less pronounced, perhaps for the reasons outlined in the introduction). It, therefore, appears more likely that it is the choice of the RB category structure employed with which to compare the II structure that is driving the qualitatively different pattern of results between studies. Nomura et al. used a unidimensional category structure while we used a conjunctive structure.

Although both conjunctive and unidimensional structures effectively manipulate the relative verbalizability of the optimal decision bound (Edmunds et al., 2015), the advantage of using the conjunctive structure is that it controls for extraneous differences that are present between the unidimensional and the II structure that have been shown in previous work to

have an important impact on categorization (e.g., Edmunds et al., 2015; Stanton & Nosofsky, 2007; Wills et al., 2013). Specifically, the unidimensional structure has only one relevant dimension but the II structure has two relevant dimensions. This means that selective attention is necessary in the RB condition but would be detrimental in the II condition (Nosofsky & Kruschke, 2002). In addition, multidimensional categorizations are typically more complex and require greater levels of cognitive resources than unidimensional categorizations (Milton et al., 2008; Pothos & Close, 2008; Wills et al., 2013, 2015). This is likely to lead to greater activation more generally in the II condition than the RB condition and perhaps particularly in the basal ganglia which has been argued to be involved in the learning of more complex category structures (e.g., Ell et al., 2010; Filoteo et al., 2005a).

Related to this, because unidimensional classifications are generally easier to learn than multidimensional classifications (e.g., Ashby, Maddox, & Bohil, 2002; Maddox, Ashby, & Bohil, 2003), Nomura et al. reduced the category separation of the unidimensional structure compared to the II structure. While this enabled error rates to be successfully matched between conditions, this manipulation effectively replaced one confound with another (Stanton & Nosofsky, 2007). The MTL is assumed to be critical for storing the precise location of the decision bound (Nomura & Reber, 2008) and it seems plausible that this would be more demanding in Nomura et al.'s unidimensional structure, where the decision boundary is more difficult to perceptually discriminate than the II structure which could have been driving the differential activation in this region.

While it may be possible to question our interpretation of these differences, our general point - that controlling between category structures for extraneous factors that have been shown to have a strong influence on categorization allows stronger inferences to be drawn - appears relatively uncontroversial. One might, of course, respond to this by arguing that behavioral dissociations predicted by COVIS have also emerged when using the same

category structures as we employed in the present study (e.g., Filoteo et al., 2010; Zeithamova & Maddox, 2006). However, these behavioral dissociations have already come under detailed critique (e.g., Newell et al., 2013; Newell et al., 2010); in this regard, our results extend these concerns to previous imaging data. While COVIS has undeniably had a positive impact on the field of categorization by motivating new lines of research and is groundbreaking in terms of the precise neurobiological predictions it makes, our results indicate that it may be in need of revision to accommodate the greater level of MTL activation in II categorization compared to RB categorization that we observed.

Another notable feature of our results is the extensive MTL activation found in the RB and, in particular, the II condition. The precise role that the MTL plays in category learning has been contentious. Some research implicates this region in RB or explicit learning alone (e.g., Nomura et al., 2007; Poldrack et al., 2001), other research shows that the MTL can also be found during II-like learning (Milton & Pothos, 2011; Cincotta & Seger, 2007), yet further research shows the hippocampus decreases in activation after initial category learning (Seger and Cincotta, 2006), while other studies found no activation at all (e.g., Seger & Cincotta, 2002; Lopez-Paniagua & Seger, 2011; Milton et al., 2009; Tracy et al., 2003). While these discrepancies may, of course, relate to the very different categorization tasks used, another possible reason is the choice of the baseline task. It is well established that a resting baseline (such as viewing a fixation cross or a blank screen) leads to activation of the default network which is known to engage the MTL (Buckner et al., 2008). It is, therefore, possible that the frequent choice of a resting baseline in categorization studies (e.g., Milton et al., 2009; Seger & Cincotta, 2002, 2005; Tracy et al., 2003) may have led to an underestimation of the involvement of the MTL in some past category learning studies.

Having said this, several theories have been proposed with regards the function of the MTL in category learning. For example, Seger et al. (2011) suggested that the anterior

hippocampus is necessary for encoding the relationship between a stimulus and a particular response (see also Chua et al., 2007) while the posterior hippocampus is required for the retrieval of the context in which a stimulus has previously been encountered (e.g., retrieving the stimulus-response mapping). Similar to this, Love and Gureckis (2007) emphasized that the MTL (and specifically the hippocampus) is particularly important for forming abstract codes (known as clusters) which represent stimulus configurations (Davis, Love & Preston, 2012a, Staresina & Davachi, 2009). New stimuli that are similar to previously seen configurations will be 'captured' by a pre-existing cluster, but if a stimulus is sufficiently novel, the MTL creates a new cluster for it (Love & Gureckis, 2007). Inspired by this account, Davis et al. (2012a; for a related study see also Davis et al., 2012b) conducted an experiment in which the stimuli (schematic beetles) could typically be classified by a single-property rule but in which there were a few stimuli that were exceptions to that rule. Davis et al. (2012a) found that the MTL was more activated for these exception stimuli than the rule consistent items, and hypothesised that this was due to it being involved in the creation of new clusters to represent the exceptions. Our results seem entirely consistent with these views. Specifically, the MTL activation observed in both the II and RB conditions may reflect that both category structures require the formation of clusters and perhaps in particular the need to encode and retrieve the category label with which a stimulus was associated. The greater MTL activation we observed in the II condition than the RB condition may reflect that the II structure is likely to evoke a greater number of exceptions to the applied rule, requiring the creation of additional clusters. The greater posterior parahippocampal gyrus activation in the second half of training compared to in the first half of training would also be consistent with the idea that the number of clusters increases as exceptions to any rule employed increase.

Our results are consistent with the growing body of evidence that the basal ganglia plays an important role in category learning (for a review see Seger, 2008). Previous work has found that the caudate head is important for the processing of positive feedback (Seger & Cincotta, 2002; Cincotta & Seger, 2007; Filoteo et al., 2005). Our finding of activation in the head of the caudate for both RB and II category learning during correct categorization trials compared with incorrect trials further underscores the important role this region has in the processing of positive feedback (Cincotta & Seger, 2007; Filoteo et al., 2005b; Seger & Cincotta, 2002). The body and tail of the caudate, as well as being linked to COVIS's implicit system, have been shown to activate more for good learners than poor learners in a rule learning task (Seger & Cincotta, 2006). Additionally, Lopez-Paniagua and Seger (2011) linked the body and tail of the caudate to stimulus-response processing. Similarly, the activation in the putamen we observed for RB and II learning may reflect motor planning demands (Cincotta & Seger, 2007). While there are clearly differences in the procedures used in these studies and ours, it is plausible that these regions serve the same role in RB and II learning. If this is the case then our results would also be consistent with Nosofsky and Stanton's (2005) claim that RB categorization as well as II categorization has a procedural component.

We also found greater activation in the medial prefrontal cortex for correct compared to incorrect responses in both the RB and II conditions. This finding is analogous to the results of Schnyer et al. (2009) who found that patients with ventromedial prefrontal cortex (VMPFC) lesions had impaired learning for both RB and II tasks compared to controls. Schnyer et al. suggested that the VMPFC is responsible for feedback processing in both RB and II learning and is involved in the selection and maintenance of the optimal learning strategies. Our imaging results therefore provide converging support for this previous neuropsychological evidence.

While fMRI provides excellent spatial resolution, it is well known to have limited temporal resolution. One consequence of this is that our study, like Nomura et al.'s (2007), and virtually all extant imaging studies of categorization cannot determine whether the activation identified is driven during the response or feedback processing stages. While the greater activation in the MTL for the II than the RB condition appears unexpected from COVIS's perspective regardless of when it occurs in the category learning process it would, nonetheless, be valuable in follow-up studies to understand at what stage in the process this difference is occurring. One possibility might be to have a subset of trials where no feedback is provided to examine the relative activation of feedback vs no feedback trials. Another option would be to include an extra variable ITI after the response has been made to identify activation differences between the response and feedback stages (see Lopez-Paniagua & Seger, 2011, for an example of where this has been done). Both approaches have challenges - for instance, in no feedback trials participants may self-generate internal feedback, particularly when they have acquired a strong understanding of the category structure. Equally, one consequence of adding an extra ITI is that it would increase the delay between making a response and receiving feedback which has been suggested to disrupt learning in COVIS's implicit system (Maddox, Ashby & Bohil, 2003; Maddox & Ing, 2005). Nevertheless, exploring this issue appears a fruitful area for future research.

Another limitation of fMRI is that it is not possible to establish whether all of the diverse areas activated are necessary for the learning of the II and RB structures. For example, it is possible that the activation in the MTL was not essential for the category learning that occurred. While this is plausible, it is still difficult though to explain from this perspective why the MTL activates more for the II condition than the RB condition; in contrast, this difference is readily compatible with the idea that II learning requires greater memory demands to compensate for the absence of an easily verbalizable rule. Nevertheless,

to test our explanation for this result, it would be valuable in future to investigate patients with MTL lesions to see if they, as we would predict from our theory, perform worse in acquiring II categories than RB categories. A further prediction derivable from our hypothesis is that people should have enhanced memory for instances after II category learning than after RB learning. As far as we are aware, the former hypothesis has not yet been investigated; however, the latter hypothesis is the subject of ongoing behavioral work in our lab.

Finally, there is a temptation to consider our data in the context of whether it is more supportive of single-system or dual-system accounts. For instance, the extensive overlap of activation between the RB and II conditions is consistent with the idea that these category structures are learned by the same neural system. According to this view, the greater activation in the II condition relative to the RB condition may just reflect that participants in the II condition, who were learning a more difficult to verbalize decision boundary, had to recruit greater neural resources to reach the same level of performance. Of course, an alternative way of explaining our results is in terms of a dual-process model such as ATRIUM (Erickson & Kruschke, 1998). In ATRIUM, one system is rule-based and is conceptually similar to COVIS's verbal system. The other system is also assumed to be explicit but is exemplar-based and is responsible for learning when rules are not easily applicable. One possibility, therefore, is that participants were utilizing a sub-optimal rule but were supplementing this with an exemplar-based process for the items in a region of decision space which did not fit into the rules that they were utilizing. The results of Davis et al. (2012a), who found that the MTL was more engaged for exception items than those stimuli which followed a simple rule, would be consistent with this explanation. While our results could, therefore, be conceptualized in either of these ways, we would generally concur with the view of Davis et al. (2012a) who note that given that the criteria for establishing truly

qualitatively separable systems are often underspecified, a more profitable way of viewing category learning may be to link brain function to particular processes required. For instance, the prefrontal cortex may be involved in hypothesis generation and rule selection, the caudate head in feedback processing, the caudate body and tail for stimulus-response associations, and the MTL in storing decision bounds and/or memory for particular exemplars. In this latter example, as discussed previously, the MTL may be an important region for category learning in general but its role could have a greater emphasis in II learning where there are less verbalizable rules than in RB learning which may encourage more specific storage of exemplars (Nosofsky et al., 2012) to supplement any rules that are applied.

Conclusion

The present study aimed to build on the limited amount of research that has directly compared the neural regions involved in RB and II category learning. We found that when we controlled for category separation, number of relevant dimensions, and error rates, extensive neural overlap in the learning of RB and II categories emerged and there was no evidence for the pattern of results predicted by COVIS. In particular, we found increased activation in the MTL, long considered critical for explicit memory (e.g., Scoville & Milner, 1957; Squire, Stark & Clark, 2004), for the II condition, which is assumed by COVIS to preferentially recruit the implicit system, compared to the RB condition. Our findings, therefore, extend our understanding of the neural processes that underlie RB and II learning and pose a challenge for COVIS as it is currently instantiated.

Acknowledgements

The support of a South West Doctoral Training Centre (SWDTC) Economic and Social Research Council (ESRC) Studentship Award (ES/J50015X/1) to the first author is appreciatively acknowledged. We also thank Todd Maddox for supplying the stimuli used in this study and Greg Ashby for his comments on this work. The participation of University of Exeter student volunteers is also greatly appreciated.

References

Akaike H (1974): A new look at the statistical model identification. IEEE Transact. Automat Contr 19: 716-723.

Ashby FG (2014): Is state-trace analysis an appropriate tool for assessing the number of cognitive systems? Psychon Bull Rev 21: 935-946.

Ashby FG, Alfonse-Reese LA, Turken AU, Waldron EM (1998): A neuropsychological theory of multiple systems in category learning. Psychol Rev 105: 442–481.

Ashby FG, Ell SW, Waldron EM (2003): Procedural learning in perceptual categorization. Mem Cognit 31: 1114–1125.

Ashby FG, Gott RE (1988): Decision rules in the perception and categorization of multidimensional stimuli. J Exp Psychol Learn Mem Cogn 14: 33-53.

Ashby FG, Maddox WT (2011): Human category learning 2.0. Annal N Y Acad Sci 1224: 147-161.

Ashby FG, Maddox WT, Bohil CJ (2002): Observational versus feedback training in rule-based and information-integration category learning. Mem Cognit 30: 666-677.

Ashby FG, Valentin VV (2005): Multiple systems of perceptual category learning: Theory and cognitive tests. Handbook of categorization in cognitive science (pp. 547-572). New York USA: Elsevier.

Buckner RL, Andrews-Hanna JR, Schacter DL (2008): The brain's default network: anatomy, function and relevance to disease. Ann N Y Acad Sci 1124: 1–38.

Chua EF, Schacter DL, Rand-Giovannetti E, Sperling RA (2007): Evidence for a specific role of the anterior hippocampal region in successful associative encoding. Hippocampus 17: 1071-1080.

Cincotta CM, Seger CA (2007): Dissociation between striatal regions while learning to categorize via feedback and via observation. J Cogn Neurosci 19: 249-265.

Davis T, Love BC, Preston AR (2012a): Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. Cereb Cortex 22: 260-273.

Davis, T, Love BC, Preston AR (2012b): Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. J Exp Psychol Learn Mem Cogn 38: 821-839.

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

Dienes, Z. (2011): Bayesian versus orthodox statistics: Which side are you on? Perspect Psychol Sci 6: 274-290.

Donkin C, Newell BR, Kalish M, Dunn JC, Nosofsky RM (2015): Identifying strategy use in category learning tasks: A case for more diagnostic data and models. J Exp Psychol Learn Mem Cogn 41: 933-948.

Duncan J (2010): The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci14: 172-179.

Edmunds CER, Milton F, Wills AJ (2015): Feedback can be superior to observational training for both rule-based and information-integration category learning. Q J Exp Psychol 68: 1203-1222.

Eickhoff SB, Paus T, Caspers S, Grosbras MH, Evans A, Zilles K, Amunts K (2007): Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. Neuroimage 36: 511-521.

Ell SW, Marchant NL, Ivry RB (2011): Focal putamen lesions impair learning in rule-based, but not information-integration categorization tasks. Neuropsychologica 44: 1737-1751.

Ell SW, Weinstein A, Ivry RB (2010): Rule-based categorization deficits in focal basal ganglia lesion and Parkinson's disease patients. Neuropsychologia 48: 2974-2986.

E-Prime [computer program] (2002): http://www.pstnet.com: Psychology Software Tools. Pittsburgh, PA

Erickson MA, Kruschke JK (1998): Rules and exemplars in category learning. J Exp Psychol Gen 127: 107-140.

Filoteo JV, Lauritzen S, Maddox WT (2010): Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. Psychol Sci 21: 415-423.

Filoteo JV, Maddox WT, Salmon DP, Song DD (2005a): Information-integration category learning in patients with striatal dysfunction. Neuropsychology 19: 212-222.

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

Friston KJ (2011): JSICM@il. SPM Archives: Fri 4th Nov 2011, 14:43

Helie S, Waldschmidt JG, Ashby FG (2010): Automaticity in rule-based and information-integration categorization. Atten Percept Psychophys 72: 1013–1031.

Jeffreys, H (1961): The Theory of Probability. (3rd ed.). Oxford: Oxford University Press.

Koenig P, Smith EE, Glosser G, DeVita C, Moore P, McMillan C, Gee J, Grossman M (2005): The neural basis for novel semantic categorization. Neuroimage 24: 369–383.

Krasnow B, Tamm L, Greicius MD, Yang TT, Glover GH, Reiss AL, Menon V (2003): Comparison of fMRI activation at 3 and 1.5 T during perceptual, cognitive, and affective processing. Neuroimage 18: 813-826.

Lewandowsky S, Yang LX, Newell BR, Kalish ML (2012): Working memory does not dissociate between different perceptual categorization tasks. J Exp Psychol Learn Mem Cogn 38: 881.

Lopez-Paniagua D, Seger CA (2011): Interactions within and between corticostriatal loops during component processes of category learning. J Cogn Neurosci 23: 3068-3083.

Love BC, Gureckis TM (2007): Models in search of a brain. Cogn Affect Behav Neurosci 7: 90-108.

Maddox WT, Ashby FG (2004): Dissociating explicit and procedural-learning based systems of perceptual category learning. Behav Process 66: 309-332.

Maddox WT, Ashby FG, Bohil CJ (2003): Delayed feedback effects on rule-based and information-integration category learning. J Exp Psychol Learn Mem Cogn 29: 650-662.

Maddox WT, Ing AD (2005): Delayed feedback disrupts the procedural-learning system but not the hypothesis testing system in perceptual category learning. J Exp Psychol Learn Mem Cogn 31: 100-107.

Maddox WT, Ashby FG, Ing AD, Pickering AD (2004): Disrupting feedback processing interferes with rule-based but not information-integration category learning. Mem Cognition 32: 582-591.

Maddox WT, Bohil CJ, Ing AD (2004): Evidence for a procedural-learning based system in perceptual category learning. Psychon Bull Rev 11: 945–952.

Maddox WT, Filoteo JV, Hejl KD, Ing AD (2004): Category number impacts rule-based but not information-integration category learning: further evidence for dissociable category learning systems. J Exp Psychol Learn Mem Cogn 30: 227– 235.

Maddox, WT, Ing, AD, Lauritzen, JS (2006): Stimulus modality interacts with category structure in perceptual category learning. Percept Psychophys 68: 1176-1190.

Maldjian JA, Laurienti PJ, Burdette JB, Kraft RA (2003): An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19: 1233–1239.

Matsuki K (2014): grt: General recognition theory. r package version 0.2: Retrieved from http://CRAN.R-project.org/package=grt

Milton F, Butler CR, Benattayallah A, Zeman AZJ (2012): The neural basis of autobiographical memory deficits in transient epileptic amnesia. Neuropsychologia 50: 3528-3541.

Milton F, Longmore CA, Wills AJ (2008): Processes of overall similarity sorting in free classification. J Exp Psychol Hum Percept Perform 34: 676-692.

Milton F, Muhlert N, Butler CR, Benattayallah A, Zeman AZJ (2011): The neural correlates of everyday recognition memory. Brain Cogn 76: 369–381.

Milton F, Pothos EM (2011): Category structure and the two learning systems of COVIS. Eur J Neurosci 34: 1326-1336.

Milton F, Wills AJ, Hodgson TL (2009): The neural basis of overall similarity and single-dimension sorting. Neuroimage 46: 319–326.

Newell BR, Dunn JC, Kalish M (2010): The dimensionality of perceptual category learning: A state-trace analysis. Mem Cognit 38: 563-581.

Newell BR, Dunn JC, Kalish M (2011): 6 Systems of Category Learning: Fact or Fantasy? Psychol Learn Motiv 54: 167-215.

Newell BR, Moore CP, Wills AJ, Milton F (2013): Reinstating the Frontal Lobes? Having More Time to Think Improves Implicit Perceptual Categorization A Comment on Filoteo, Lauritzen, and Maddox (2010). Psychol Sci 24: 386-389.

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005): Valid conjunction inference with the minimum statistic. Neuroimage 25: 653-660.

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

Nomura EM, Reber PJ (2008): A review of medial temporal lobe and caudate contributions to visual category learning. Neurosci Biobehav Rev 32: 279-291.

Nosofsky RA, Kruschke JK (2002): Single system models and interference in category learning: Commentary on Waldron & Ashby (2001). Psychon Bull Rev 9: 169–174.

Nosofsky RM, Little DR, James TW (2012): Activation in the neural network responsible for categorization and recognition reflects parameter changes. Proc Nat Acad Sci U.S.A. 109: 333-338.

Nosofsky RM, Palmeri TJ, McKinley SC (1994): Rule-plus-exception model of classification learning. Psychol Rev 101: 53–79.

Nosofsky RM, Stanton RD (2005): Speeded classification in a probabilistic category structure: contrasting exemplar-retrieval, decision-boundary, and prototype models. J Exp Psychol Hum Percept Perform 31: 608-629.

Poldrack RA, Clark J, Pare-Blagoev EJ, Shohamy D, Moyano JC, Myers C, Gluck MA (2001): Interactive memory systems in the human brain. Nature 414: 546-550.

Poldrack RA, Foerde K (2008): Category learning and the memory systems debate. Neurosci Biobehav Rev 32: 197-205.

Pothos EM, Close J (2008): One or two dimensions in spontaneous classification: a simplicity approach. Cognition 107: 581–602.

Reber PL, Stark CEL, Squire LR (1998): Cortical areas supporting category learning identified using functional magnetic resonance imaging. Proc Natl Acad Sci USA 95: 747–750.

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

Scoville WB, Milner B (1957): Loss of recent memory after bilateral hippocampal lesions. J Neurol Neurosurg Psychiatry 20: 11-21.

Seger CA (2008): How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. Neurosci Biobehav Rev 32: 265–78.

Seger CA, Cincotta CM (2002): Striatal activation in concept learning. Cogn Affect Behav Neurosci 2: 149-161.

Seger CA, Cincotta CM (2005): The roles of the caudate nucleus in human classification learning. J Neurosci 25: 2941-2951.

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

Seger CA, Dennison CS, Lopez-Paniagua D, Peterson EJ, Roark AA (2011): Dissociating hippocampal and basal ganglia contributions to category learning using stimulus novelty and subjective judgments. Neuroimage 55: 1739-1753.

Smith JD, Boomer J, Zakrzewski AC, Roeder JL, Church BA, & Ashby FG (2014): Deferred Feedback Sharply Dissociates Implicit and Explicit Category Learning. Psychol Sci 25: 447–457.

Song XW, Dong ZY, Long XY, Li SF, Zuo XN, Zhu CZ, He Y, Yan CG, Zang YF (2011): REST: a toolkit for resting-state functional magnetic resonance imaging data processing. PloS one 6: e25031.

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

Stanton RD, Nosofsky RM (2007): Feedback interference and dissociations of classification: Evidence against the multiple-learning-systems hypothesis. Mem Cognit 35: 1747-1758.

Stanton RD, Nosofsky RM (2013): Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. J Exp Psychol Learn Mem Cogn 39: 1174-1191.

Staresina BP, Davachi L (2009): Mind the gap: binding experience across space and time in the human hippocampus. Neuron 63: 267-276.

Stark CE, Squire LR (2001): When zero is not zero: the problem of ambiguous baseline conditions in fMRI. Proc Nat Acad Sci U.S.A. 98: 12760-12766

Squire LR, Stark CE, Clark RE (2004): The medial temporal lobe. Annu Rev Neurosci 27: 279-306.

Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging.

Tracy JI, Mohamed F, Faro S, Pinus A, Tiver R, Harvan J, Bloomer C, Pyrros A, Madi S (2003): Differential brain responses when applying criterion attribute versus family resemblance rule learning. Brain Cogn 51: 276–286.

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

Waldschmidt JG, Ashby FG (2011): Cortical and striatal contributions to automaticity in information-integration categorization. Neuroimage 56: 1791-1802.

Wills AJ, Inkster AB, Milton F (2015): Combination or differentiation? Two theories of processing order in classification. Cogn Psych 80: 1-33.

Wills AJ, Milton F, Longmore CA, Hester S, Robinson J (2013): Is overall similarity classification less effortful than single-dimension classification? Q J Exp Psychol 66: 299-318.

Worthy DA, Markman AB, Maddox W (2013): Feedback and stimulus-offset timing effects in perceptual category learning. Brain Cogn 81: 283-293.

Xie Z, Maddox WT, McGeary JE, Chandrasekaran B (2015): The C957T polymorphism in the dopamine receptor D2 (DRD2) gene modulates domain-general category learning. J Neurophysiol 113: 3281-3290.

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

Zink CF, Pagnoni G, Chappelow J, Martin-Skurski M, Berns GS (2006): Human striatal activation reflects degree of stimulus saliency. Neuroimage 29: 977-983.

Footnotes

[1] Bayes Factor analysis requires an estimate of the mean expected difference under the experimental hypothesis; we estimated this from Filoteo et al's (2010) study, which used the same stimuli and category structures, using plot digitizer (https://sourceforge.net/projects/plotdigitizer/). Following Dienes (2011), the expected difference was modeled as a two-tailed normal distribution with a standard deviation equal to half the mean. By convention, a Bayes factor of over three is interpreted as providing substantial evidence for the experimental hypothesis (Jeffreys, 1961), while a Bayes factor below a third provides substantial evidence for the null (Dienes, 2011). A value in between a third and three is indeterminate, providing no clear evidence either for the null or the experimental hypothesis.

[2] The percent signal change in the right caudate body for the RB and II conditions in Nomura et al.'s (2007) study (shown in their Figure 4d) was used to calculate the prior (these values were estimated using plot digitizer https://sourceforge.net/projects/plotdigitizer/). The expected difference was modeled as a two-tailed normal distribution with a standard deviation equal to half the mean (Dienes, 2011).

Figure Legends

Figure 1. Examples of unidimensional, conjunction, and information-integration category structures. Each open circle represents one member of category A; each filled square represents one member of category B. Figure adapted from Wills et al. (2013) and Zeithamova and Maddox (2006).

Figure 2. The category structures used for the present study (a) The conjunctive rule-based condition; (b) The information-integration condition. Solid lines indicate the decision boundary separating category A (unfilled circles) and category B (filled squares).

Figure 3. Mean performance across runs in the RB and II conditions. Error bars show standard error.

Figure 4. Whole brain analyses on all runs of the study for: (a) areas of activation in the RB condition; (b) areas of activation in the II condition (c) a conjunction analysis showing areas commonly activated in the RB and II conditions. All analyses are thresholded at $p<.001$ and 27 contiguous voxels. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Figure 5. Analyses of areas activated in all runs of the study (a) Whole-brain analysis of areas more active in the II condition compared with the RB condition; (b) Regions of the MTL more active in the II condition compared with the RB condition; non-MTL regions

were masked in this analysis but the thresholds remained p<.001 and 27 contiguous voxels. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Figure 6. Analysis of blocks 3 and 4 for: (a) Areas more activated in the II condition than the RB condition. (b) Regions of the MTL more active in the II condition compared with the RB condition; non-MTL regions were masked in this analysis but the thresholds remained at p<.001 and 27 contiguous voxels. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Figure 7. Analysis of participants who were shown by the modeling analysis to use the optimal learning strategy overall for all runs of the study: (a) Areas commonly activated in the RB condition and the II condition (with thresholds of p<.001 and 27 contiguous voxels); (b) A ROI analysis of areas of the MTL more activated in the II condition compared with the RB condition (with thresholds of p<.05 and 79 contiguous voxels). The right most image represents the brain from the bottom. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Supplementary Figure Legends

Supplementary Figure 1. Whole brain analysis of all runs comparing incorrect trials to the 'odd or even' baseline, showing areas of common activation in the II and RB conditions thresholded at $p<.001$ and 27 contiguous voxels. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Supplementary Figure 2. Analysis of areas commonly activated in both the RB and II conditions in runs 3 and 4 of the study only thresholded at $p<.001$ and 27 contiguous voxels. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Supplementary Figure 3. Analysis of correct trials contrasted against incorrect trials across all participants showing a ROI analysis of MTL activation greater in the II compared with the RB condition in runs 3 and 4 (thresholded at $p<.05$ and 79 contiguous voxels). Non-MTL regions are masked. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.