The neural correlates of similarity- and rule-based generalization.

Fraser Milton[1], Pippa Bealing[1], Kathryn L. Carpenter[1], Abdelmalek Bennattayallah[2]

& Andy J. Wills[3]

[1] Discipline of Psychology, University of Exeter, Exeter, U.K.
[2] Exeter Medical School, University of Exeter, UK
[3] School of Psychology, University of Plymouth, Plymouth, U.K.

Address for correspondence:
Fraser Milton
Washington Singer Laboratories,
Perry Road,
Exeter
U.K.
EX4 4QG
Tel: +44 1392 725884

E-mail: f.n.milton@exeter.ac.uk

Abstract

The idea that there are multiple learning systems has become increasingly influential in recent years with many studies providing evidence that there is both a quick, similarity, or feature-based, system, and a more effortful, rule-based system. A smaller number of imaging studies have also examined whether neurally dissociable learning systems are detectable. We further investigate this by employing for the first time in an imaging study a combined positive and negative patterning procedure originally developed by Shanks and Darby (1998). Unlike previous related studies employing other procedures, rule generalization in the Shanks-Darby task is beyond any simple non-rule-based (e.g., associative) account. We found that rule- and similarity-based generalization evoked common activation in diverse regions including the prefrontal cortex and the bilateral parietal and occipital lobes indicating that both strategies likely share a range of common processes. No differences between strategies were identified in whole-brain comparisons but exploratory analyses indicated that rule-based generalization led to greater activation in the right middle frontal cortex than similarity-based generalization. Conversely, the similarity group activated the anterior medial frontal lobe and right inferior parietal lobes more than the rule group did. The implications of these results are discussed.

Key words: rules, similarity, categorization, generalization, fMRI

The ability to generalize information we have previously learned to novel stimuli is fundamental for successful functioning in our everyday environment. An enduring and contentious question is whether this is achieved by separable learning systems (e.g., Ashby et al., 1998; Brooks, 1978) or just a single system (e.g., Newell et al., 2011; Nosofsky & Kruschke, 2002). Multiple-systems accounts typically posit the existence of a non-deliberative (Wills et al., 2013) or non-analytic (Brooks, 1978) process, that is automatic (Smith et al., 1998), similarity-based (Milton, Longmore, & Wills, 2008), and driven by associative (McLaren, Green, & Mackintosh, 1994) or implicit (Ashby et al., 1998) processes. A second system is assumed to be deliberative (Wills et al., 2013) or analytic (Brooks, 1978), controlled (Smith et al., 1998), rule-based (Ashby et al., 1998), and requiring of extensive cognitive resources (Wills et al., 2015). In this article, we refer to these two systems as similarity and rule-based.

Much of the evidence relevant to this debate has come from behavioral or comparative studies. Some of this evidence is consistent with multiple learning systems accounts (e.g., Allen & Brooks, 1991; Ashby & Maddox, 2011; Kemler Nelson, 1984; Maes et al., 2015; Rips, 1989), while others maintain this evidence can be more parsimoniously explained by a single system (e.g., Edmunds et al., 2015; Newell et al., 2013; Stanton & Nosofsky, 2013; Wills, Inkster, & Milton, 2015). Consequently, there is currently no clear consensus on this issue. A complimentary, and currently relatively underexplored, approach is to use brain imaging to examine whether there are neurally dissociable learning systems. One such fMRI study, loosely based on earlier behavioral work by Allen and Brooks (1991), was conducted by Koenig et al. (2005) who asked participants to classify a set of cartoon animals differing on four stimulus dimensions (e.g., legs, neck type). Participants in the rule condition were informed of a complex rule (category membership requires the

instance to possess three out of four characteristic features for that category). In the similarity condition participants were not told the rule but instead asked to make a quick decision using their first impressions about which category a particular instance was more similar to. Both groups were provided with trial-by-trial feedback. Koenig et al. found that similarity, compared to rule-based, categorization recruited greater activation in bilateral temporo-parietal regions as well as bilateral anterior prefrontal regions (BA 10). Conversely, the rule-based condition led to greater activation than the similarity condition in the left frontal lobes, left inferior parietal lobes, and the right superior parietal lobes. One feature of this study, however, is that it is not clear exactly what strategy participants in the similarity condition are employing, which makes interpretation of the imaging results more complicated. Specifically, Koenig et al. assume that participants are using a similarity-based approach that presumably requires the use of most, if not all, of the dimensions. While this is plausible, an alternative explanation is that participants in the similarity condition are using a simpler rule-based approach, such as a single dimension-plus-exception strategy (e.g., Ward & Scott, 1987) which could also result in the level of performance obtained. In this latter case, participants in the similarity condition are using *fewer* of the dimensions than those in the rule condition.

In a study somewhat more closely based on the work of Allen & Brooks (1991), Patalano et al. (2001) also observed activation in bilateral frontal cortex in the rule-based condition that was not present in the similarity condition. Occipital lobe and cerebellum activation was prevalent in both conditions. However, significant neural differences between the groups were relatively restricted, even though one-tailed tests were used. For example the greater frontal lobe activation in the rule than the similarity condition was only marginally significant ($p = .06$).

Using a slightly different approach - the criterial attribute procedure, based on earlier behavioral work by Kemler Nelson (1984) - Tracy et al. (2003) investigated the neural correlates of family resemblance categorization (assumed to use the similarity system) and unidimensional categorization (assumed to employ the rule system). Similar to the category structure employed by Koenig et al. (2005), a family resemblance category (e.g., Rosch & Mervis, 1975) possessed a number of characteristic but not defining features – an item did not have to possess any single feature or features as long as it possessed enough characteristic features (3 out of 4 typical features) of that category. In contrast, a unidimensional category was based around a single defining feature that the authors assumed required use of the rule-based system. Tracy et al. (2003) found greater activation in the extrastriate cortex (BA's 18 and 19) and the left cerebellum for family resemblance (similarity-based) categorization than unidimensional categorization, while unidimensional categorization led to greater activation in bilateral frontal lobes than family resemblance categorization. However, recent behavioral model-based analysis suggests that family resemblance categorization in the criterial attribute procedure is often due to the use of a single non-criterial dimension, which is a strategy not detectable by the standard analysis employed by Tracy et al. (for a detailed discussion, see Wills et al., 2015). This again makes interpretation of the neural differences observed more difficult.

In contrast to Tracy et al.'s (2003) conclusions, Milton, Wills, and Hodgson (2009) proposed that both family resemblance and unidimensional categorization are the result of a single rule-based system, with family resemblance categorization requiring a more complex, multi-dimensional, rule than unidimensional categorization (see also Wills et al., 2013). Consistent with their proposal, Milton et al. found

extensive common activation between family resemblance and unidimensional categorization including the dorsolateral frontal cortex and the anterior cingulate. The most notable difference between groups was the greater right ventrolateral frontal cortex activation for family resemblance than unidimensional categorization, which the authors proposed indicated the greater working memory resources required to employ a multidimensional rule.

A different approach was taken by Nomura et al. (2007), who conducted an fMRI study based on the influential COVIS framework (Ashby et al., 1998); participants viewed a series of Gabor patches and learned either a rule-based (RB) task that possessed an easily verbalizable, unidimensional rule ("thinner lines belong in category A, thicker lines in category B") which is assumed to encourage use of the explicit system or an information-integration (II) task which requires participants to combine information from two unrelated stimulus dimensions. The optimal II category structure is assumed to be difficult or impossible to verbalize, which should encourage use of COVIS's implicit system. In line with COVIS's predictions, dissociable neural activation was found with the medial temporal lobes (MTL) more activated in RB compared with II learning, and the caudate body more engaged in II than RB learning.

While intriguing, 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) was smaller in the RB than the II condition and the selective attention demands were greater in the RB than the II condition (as only one of the two dimensions was relevant to learn the RB structure while both were required for the II structure) meaning that non-essential differences could have been driving the neural dissociations. In a recent study conducted by Carpenter, Wills,

Benattayallah, and Milton (in press) when these non-essential differences between the RB and II conditions were better equated (by comparing a conjunctive RB structure against a standard II structure), the pattern of results observed by Nomura et al. (2007) did not emerge and instead there was extensive common overlap between the conditions. Furthermore, the II condition evoked greater activation in the MTL than the RB condition, which may reflect the greater memory demands in the II condition where no rule was readily available. In another related study, albeit one which used very different stimuli (the stimuli varied on rectangle height and width of an ellipse), Milton and Pothos (2011) compared activation between a RB task and an II-like task but found minimal neural dissociations and instead found extensive overlap of activation suggesting that both groups were using similar neural processes.

Finally, Grossman et al. (2002), using a modified version of Rips's (1989) classic procedure gave participants a description of an item such as "a round object 2 inches in diameter" who had to assign it to either the category of "quarter" or "pizza". The description is more similar in size to a quarter than a pizza but a pizza has a variable diameter (so could, in principle, be 2 inches) while a quarter does not (so it cannot be 2 inches). Participants who choose the quarter category were assumed to be making a similarity judgment while those who assign it to the pizza category are following a rule. Grossman et al. found that there was greater recruitment of the left dorsolateral prefrontal cortex for rule than similarity responses while the right inferior parietal lobe, which they noted is involved in overall feature configuration (Wilkinson et al., 2002), was activated more for similarity- than rule-based responses. However, Nosofsky and Johansen (2000) have demonstrated that the results from this procedure can be accommodated by a simple, single-process, exemplar-based learning system without requiring qualitatively distinct systems for the different strategies.

We investigate whether there are neurally separable rule and similarity generalization systems from a different angle using a procedure, based on Shanks and Darby's (1998) Experiment 2, which has not previously been examined using brain imaging. The design of this experiment is shown in Figure 1. Participants took the role of an allergist who had to determine whether the meals a hypothetical patient, Mr X, eats will cause an allergic reaction or not. Letters in Figure 1 stand for particular foods (e.g., pasta or eggs), + indicates that an allergic reaction will develop and - that no allergic reaction will occur. During training, participants learn two complete negative patterning problems (e.g., A+, B+, AB-) and two complete positive patterning problems (e.g., C-, D-, CD+). Critically, however, there are also four incomplete patterning problems - for example, participants are trained on I+ and J+ but not on the outcome of I and J combined and trained that eating KL together leads to an allergic reaction but not what happens when K and L are eaten separately. During the test phase, as well as being tested on items they studied during training (e.g., I+, J+ and KL+), participants have to generalize the knowledge they have obtained to what will henceforth be referred to as the *critical* items (e.g., IJ, K, and L; shown in bold in Figure 1) and are provided no feedback on their responses.

In the case of IJ, if participants are using a similarity-based strategy then they should predict an allergic reaction as it is similar to I and J, both of which lead to an allergic reaction alone. Equally, when presented with K or L alone they should predict an allergic reaction because they are similar to KL which results in an allergic reaction. In contrast, if participants have learned the "opposites" rule from training - single foods predict the opposite to their compounds - they can use this to generalize to novel items. In this case, IJ should lead to no allergic reaction because it is the opposite outcome to I or J when presented alone. Similarly, K or L, when presented

separately, should lead to no allergic reaction as this is the opposite to KL which resulted in an allergic reaction. Shanks and Darby (1998) found that participants with high accuracy during training produced more rule-based responses for these critical test items than participants with lower accuracy. They explained this by postulating that there is a transition from a similarity to a rule-based approach, which can only be used when the basic associations have been acquired (see also Wills et al., 2011).

While using a novel procedure to compare the neural correlates of the purported rule and similarity systems is of value in itself, the Shanks-Darby procedure has some particular advantages that make it well equipped to provide new insight into this debate. First, both similarity and rule-based responses require utilizing the same number of stimulus dimensions. This is in contrast to many of the studies described above (e.g., Milton & Pothos, 2011; Milton et al., 2009; Nomura et al., 2007; Tracy et al., 2003) where the number of stimulus dimensions utilized in the similarity condition seem unlikely to be the same as in the rule conditions (it may either be more, as is commonly assumed, or sometimes less, depending on how participants approach the task in the similarity condition). Across a range of different procedures, categorizing by a larger number of dimensions is more effortful than categorizing by fewer dimensions (e.g., Edmunds et al., 2015; Milton & Wills, 2004; Wills et al., 2015). It is plausible that this could be driving the difference in neural activation between the groups, rather than indicating the involvement of qualitatively different systems. A further advantage of employing the Shanks-Darby procedure is that a full explanation of the "opposites" rule generalization is commonly thought to be beyond published associative accounts (see Maes et al., 2015 for a further discussion) allowing clear inferences to be drawn. This does not appear to be the case for the other procedures described above. For instance, non-rule based accounts, such as

Kruschke's (1992) ALCOVE model, which have a mechanism for dimensional attention are able to account for the purported rule-based classification in these tasks.

In summary, then, the present study uses a novel approach to investigate the neural differences between rule-based and similarity-based generalization. We predicted, based on previous behavioral and comparative work with this procedure (Maes et al., 2015; Wills et al., 2011), that we would observe neural differences between the generalization strategies. In particular, we hypothesized that there would be greater frontal lobe activation in the rule-based condition than the similarity-based condition (e.g., Grossman et al., 2002; Milton et al., 2009; Nomura & Reber, 2008; Patalano et al., 2001). Our prediction for which regions would be implicated in similarity-based generalization was more tentative given the greater heterogeneity in previous studies but viable options a priori included the right inferior parietal lobes (Grossman et al., 2002), and the occipital lobes (e.g., Nomura et al., 2007; Patalano et al., 2001)/ extrastriate cortex (Tracy et al., 2003).

Method

Participants

62 right-handed participants were recruited from the University of Exeter participant pool. Participants were either volunteers, received course credits, or were paid £7. Participants all gave informed consent according to procedures approved by the Psychology Ethics Committee, University of Exeter. A learning criterion was set as significantly above chance accuracy in the second half of training to ensure that all participants included in the analyses had clear evidence of learning. Without this, one could not reasonably expect true generalization to occur. This resulted in the exclusion of 10 participants. A further 14 participants did not show clear evidence of either

rule- or similarity-based generalization - defined as significantly above chance (64.6%) strategy-consistent responding for the critical test trials. These participants who did not adopt a clear strategy would likely obscure any differences that emerged between participants who did demonstrate clear rule or similarity generalization so we consequently excluded them from all the principal analyses. We, however, consider the test phase data for these 14 participants who used a mixture of rule and similarity consistent responses separately. This left 38 participants in total for our principal analyses; 24 rule-based responders and 14 similarity-based responders. This trend for a greater proportion of rule responders was not significant, $\chi^2(1) = 2.632$, $p = .105$.

Stimuli

The stimuli (food names) were identical to those used in Experiment 2 of Shanks and Darby (1998). For half of the participants the food names A-P (see Figure 1) were cheese, garlic, milk, mushrooms, seafood, red meat, olive oil, coffee, banana, eggs, orange squash, bread, avocado, peanuts, pasta, and chocolate. For the other half, the foods assigned to A/B were exchanged for those assigned to C/D and likewise for E/F and G/H, I/J and K/L, and M/N and O/P.

Procedure

Prior to entering the scanner, participants were asked to take the role of a food allergist and to learn when Mr. X would develop an allergic reaction after eating a meal containing certain foods. 29 participants were additionally provided with instructions outlining the rule (e.g., “If Mr. X is allergic to a food when it is presented on its own, he won't be allergic to it when it is presented together with another food. Conversely, if Mr X is not allergic to a food when it is presented on its own then he

will be if it is presented in combination with another food") and 33 participants were provided with instructions designed to encourage a similarity-based approach ("When making a response, please use your intuition as to what you feel is the correct answer based on what you have previously seen"). The rationale behind this was to facilitate obtaining a sufficient number of participants who consistently categorized the critical trials by either a similarity or a rule-based approach rather than to look at the neural effects of differing instructions per se. In practice, however, this instructional manipulation had no significant impact on the strategy used (we suspect, in hindsight, that it would have been more effective if, like the training items, it had been presented inside the scanner in the same context as where learning took place) and will, consequently, not be discussed further.

Visual 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. Button-press responses and reaction times were measured using a fiber-optic button box. E-Prime (Psychological Software Tools, 2002, http://www.pstnet.com) was used for the presentation and timing of stimuli and collection of response data.

In the training phase, participants received six blocks of trials, divided into two scanning runs of three blocks, with each of the 18 training stimuli (see Figure 1) presented twice in each block in a random order. Each trial began with a white screen lasting a random interval between 500-4000ms, before a black fixation cross was presented in the middle of the screen for 250ms. A meal, food names presented in black font, was then presented in the middle of a white screen for 3000ms during which time participants indicated whether it would lead to an allergic reaction (by pressing the left button box key) or would not lead to an allergic reaction (by pressing the right button box key). Following this, feedback ("Correct", in blue or "Incorrect",

in red) was presented for 500ms. For participants who failed to respond when the meal was presented, the message “Timeout!!" (in red) appeared instead. The next trial then immediately began. At the end of each block of 36 trials, the average accuracy on that block was displayed on the screen for 12 seconds, before the next block began.

The test phase included all 24 stimuli, comprising the 18 training stimuli plus the 6 critical generalization stimuli (shown in bold in Figure 1). The 18 training stimuli were presented once in each block while the 6 critical trials were each presented twice leading to 30 trials in each of the four blocks. Stimuli were presented in a random order. The intra-trial structure was identical to the training phase except that after a response a 500ms blank screen appeared rather than feedback. If no response was made, a time out message appeared as in the training phase.

fMRI Data Acquisition

Images were collected using a 1.5-T Phillips Gyroscan magnet equipped with a Sense coil. A T2*-weighted echo planar sequence was used (Tr = 3000ms, Te = 45ms, flip angle = 90˚, 32 transverse slices, FOV = 240mm, 3.5x2.5x2.5mm). The training phase comprised two runs of 240 scans and the test phase one run of 260 scans. 5 dummy scans were performed prior to the start of each stimulus sequence. Standard volumetric anatomical MRI was performed after functional scanning by using a 3-D T1-weighted pulse sequence (Tr = 25ms, Te = 4.1ms, flip angle = 30˚, 160 axial slices, 1.6x0.9x0.9mm).

Analysis of fMRI data

Analyses were carried out using SPM8 software (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 realigned T2* volumes which were spatially smoothed using a Gaussian kernel of 8mm full-width half maximum. Data were high-pass filtered (1/128 Hz) to account for low frequency drifts. The BOLD response was modeled by a canonical hemodynamic response function.

All analyses were conducted using the general linear model. In the individual participant models, the critical trials that were consistent with their overall favored strategy (i.e., rule- or similarity-based generalization) were included as one regressor, while critical trials inconsistent with this approach were a second regressor. The familiar items were partitioned into correct and incorrect responses. The duration of each event was modeled as the participant's reaction time for that trial (see Grinband et al., 2008, for the advantages of using this "variable epoch" approach). Time outs were included as a fifth regressor of no interest. The six head movement parameters were included as additional covariates. Contrasts comparing strategy-consistent responses for the critical trials were subtracted against the implicit baseline (the intervals between the five event types listed above; cf., Milton et al., 2009; Tracy et al., 2003, for a similar approach) and correct familiar trials were likewise compared to the implicit baseline. These comparisons were then included in random-effects analyses. For these analyses, participants were divided into those who provided clear evidence of either similarity or rule-based generalization (i.e., significantly above chance strategy-consistent responding on the critical generalization trials).

Whole-brain analyses were completed using a combined statistical threshold of $p<.001$ (uncorrected) and a threshold of 100 contiguous voxels, which together produce an overall corrected threshold of $p<.05$. These values were estimated using

3dClustSim as implemented in the AFNI toolbox (http://afni.nimh.nih.gov/afni/). For this, we used a smoothness estimate of 10.1x10.1x9.6 mm (this was a group level estimate calculated in SPM8 using the group residuals from the general linear model, e.g., Kiebel et al., 1999).In addition, to measure common activation between rule-based and similarity-based participants, conjunction analyses were performed. To do this, the relevant 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 contrasts were again conducted with a combined threshold of $p<.001$ (uncorrected) and a cluster threshold of 100 contiguous voxels. Note that this analysis is conservative because it reveals only those regions significantly activated for both the rule ($p<.05$, corrected) *and* the similarity ($p<.05$, corrected) conditions.

After performing the whole-brain analyses we decided to conduct more exploratory region of interest (ROI) analyses (using the WFU Pickatlas, e.g., Maldjian, Laurienti, Burdette, & Kraft, 2003) when directly comparing rule and similarity generalization. These post-hoc ROI analyses were based on our a priori predictions of regions we thought would be differentially involved between strategies and comprised the prefrontal cortex (e.g., Milton et al., 2009; Milton & Pothos, 2011), the occipital lobes/extrastriate cortex (BAs 18 and 19; e.g., Nomura et al., 2007; Tracy et al., 2003), and the right inferior parietal lobes (Grossman et al., 2002). While these exploratory analyses should accordingly be taken with some caution, we believe that they help to characterize better the nature of our results, which is particularly important given that this is the first imaging study of the Shanks-Darby procedure. For these analyses, we used thresholds of $p<.001$ and 64 contiguous voxels which together produce an overall corrected threshold of $p<.05$, as estimated by 3dClustSim.

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). [1]

Results

Behavioral analyses

Training phase

The proportions of timeouts were low in both the rule (M = .015, SD = .021) and similarity (M = .020; SD = .018) groups and there was no significant difference between them, t (31.1) = .829, p = .413. One sample t-tests revealed that the average performance in the second half of training (blocks 4-6) was significantly above chance for both the rule-based (M = .872; SD = .092; t (23) = 19.647, p <.001) and the similarity-based (M = .739; SD = .088; t (13) = 10.217, p<.001) groups, although, as in Shanks and Darby (1998), the rule group had higher accuracy than the similarity group, t (28.6) = 4.409, p < .001. Median reaction times (RT) were longer in the similarity group (1291 ms) than in the rule group (1010 ms), t(21.8) = 3.874, p < .001.

Test phase

The proportions of timeouts were again low (rule group: M = .012; SD = .016; similarity group: M = .026; SD = .031) and there was no significant difference between conditions, t(17.2) = 1.608, p = .126. Average performance for the familiar items (i.e., those seen during the training phase) across blocks is displayed in Figure 2a. The average accuracy (collapsed across blocks) for both the rule (M = .923; SD = . 083, t (23) = 25.111, p<.001) and similarity (M = .763; SD = .129; t (13) = 7.616, p <.001) groups was significantly above chance, although as in the training phase, rule-

based participants had higher accuracy than similarity-based participants, t (19.3) = 4.151, p < .001. Median RTs were non-significantly longer in the similarity group (1208 ms) than the rule group (1060 ms), t(18.4) = 1.78, p = .09.

Of particular importance, given our interest in generalization strategies, both the rule and similarity groups used their preferred strategy significantly above chance levels (rule group, M = .862; SD = .106, t (24) = 16.691, p<.001; similarity group, M = .825; SD = .095, t (13) = 12.868, p <.001) for the critical test items (see Figure 2b), and there was no significant difference in strategy-consistent responding between groups, t (29.3) = 1.074, p = .292, with substantial evidence for the null, BF = 0.25.[2] Median RTs were non-significantly shorter in the similarity group (1119 ms) than the rule group (1294 ms), t(22.3) = 1.86, p = .08.

Although participants were classified on the basis of being either rule-consistent or similarity-consistent collapsed across all critical items it does not necessarily follow that both the critical compound and element stimuli show this pattern. We therefore consider generalization to compound and element stimuli separately as in past work (e.g., Shanks & Darby, 1998; Wills et al., 2011). The mean probability of predicting an allergic reaction to the critical compound stimuli (i.e., IJ and MN) is shown in Figure 3a. As expected, there was a significant interaction between strategy used and stimulus type, F(1,36) = 382.02., p <. 001. No other main effects or interactions were significant (Ps >.25). The rule group showed rule-consistent generalization to compounds, t(23) = 25.87, p < .001, while the similarity group showed similarity-consistent generalization, t(13) = 8.21, p < .001. Median RTs were non-significantly shorter in the similarity group (1313 ms) than the rule group (1457 ms), F(1, 36) = 1.29, p = .26. No other effects were significant (Ps > .12).

Figure 3b shows the probability of predicting an allergic reaction to the critical element stimuli (i.e., K/L and O/P). There was again a significant interaction between strategy used and stimulus type, $F(1,36) = 201.80$, $p < .001$. No other main effects or interactions were significant (Ps > .4). The rule group showed rule-consistent generalization to elements, $t(23) = 11.75$, $p < .001$, while the similarity group showed similarity-consistent generalization, $t(13) = 9.43$, $p < .001$. Median RTs were non-significantly shorter in the similarity group (1123 ms) than the rule group (1289 ms), $F(1,36) = 3.96$, $p = .054$. No other main effects or interactions were significant (Ps > .06).

Imaging analyses

Training

A contrast comparing all trials and groups against the implicit baseline revealed extensive activation including diverse regions of the bilateral prefrontal cortex, bilateral parietal lobes and bilateral occipital lobes (see Figure 4a). We then compared performance on early training (Blocks 1-3) to late training (Blocks 3-6) for all trials and participants. A large cluster including the caudate head and body, which have been extensively linked to category learning (e.g., Seger, 2008), and the thalamus, was activated more early in training than later in training (peak voxel: x = 16, y = -32, z = 16). Conversely, the right inferior frontal gyrus (peak voxel: x = 24, y = 25, z = -5) and the anterior cingulate /medial prefrontal gyrus (peak voxel: x = 8, y = 30, z = 22) were activated more late in training than early in training.

For the rule group, comparing correct responses against the baseline revealed extensive activation including in the prefrontal cortex, parietal lobes, and the occipital lobes (Figure 4b). Similar regions were recruited by the similarity group (Figure 4c).

This extensive common overlap of activation was confirmed by a conjunction analysis (Figure 4d; Table 1). When directly comparing the groups, no regions were more activated by the similarity group than the rule group; however, the bilateral posterior cingulate/precuneus (cluster size: 166 voxels) was engaged more in the rule group than the similarity group (Figure 5a). An exploratory analysis of this rule -similarity contrast, with more liberal thresholds ($p<.001$, 25 contiguous voxels), and which should consequently be taken with caution, revealed two clusters in right middle frontal gyrus (1^{st} cluster, peak voxel x=32, y = 24, z = 15, cluster size: 41 voxels; 2^{nd} cluster, peak voxel, x = 28, y = 23, z = 39, cluster size: 50 voxels).

No differences emerged between groups when considering just the first half of training. When looking at the second half of training alone, there were again no areas more activated by the similarity group than the rule-based group. However, the bilateral posterior cingulate/precuneus was again activated more by the rule group than the similarity group and the right anterior cingulate/ medial frontal gyrus was also engaged (see Figure 5b).

Test

Critical trials (Generalization)

A number of brain regions were activated by similarity responders including bilateral inferior and superior parietal lobes, right middle occipital gyrus, and left medial frontal gyrus (Figure 6a). Rule-based responders engaged the left superior parietal lobes, bilateral inferior parietal lobes, bilateral middle frontal gyrus, left medial frontal gyrus, right inferior frontal gyrus and bilateral occipital lobes (Figure 6b). A conjunction analysis (see Figure 6c; Table 2) revealed extensive common overlap of activation between the similarity- and rule-based participants which included the left

superior parietal lobes, bilateral inferior parietal lobes, bilateral medial frontal gyrus, left middle frontal gyrus and the bilateral occipital gyrus.

Next, we directly contrasted brain activation between the similarity and rule groups. In contrast to the extensive common activation, no differences were identified in whole-brain analyses. However, in the exploratory ROI analyses (comprising the prefrontal cortex, right inferior parietal lobes, and bilateral occipital lobes, with thresholds of $p<.001$ and 64 contiguous voxels) we found that the right middle frontal gyrus (see Figure 7a; BA 9) was activated more for the rule group than the similarity group (in the same region as identified by the exploratory analysis documented in the rule - similarity comparison for the training phase). In contrast, we observed greater activation in the anterior medial frontal lobes (BA 10) and the right inferior parietal lobes (BA 40) for the similarity group compared to the rule-based group (Figure 7b).

Element vs compound critical stimuli

As a supplementary question, we assessed whether there were activation differences in the element (i.e., K/L and O/P) and compound (i.e., IJ and MN) critical trials. As before, only trials that were consistent with the preferred strategy of the participants (i.e., rule- or similarity-based) were included. For the rule group, there was greater activation in the occipital lobes/cerebellum and the left caudate body for the compound stimuli than for the element stimuli (see Figure 8a). In contrast, no regions were more activated for the element stimuli than the compound stimuli. For the similarity group, no areas were more active for the compound stimuli than the element stimuli, although the occipital lobes were, as for the rule-based group, activated at lower thresholds (this could reflect, in part, the smaller sample size of the similarity group compared to the rule group). However, the left precentral/postcentral gyrus was

activated more for the element stimuli than the compound stimuli (Figure 8b). When comparing the rule and similarity groups, no significant differences emerged.

Familiar items

Although not our primary focus, we also examined the brain activation for the familiar items to supplement the critical generalization trials analyses. Participants in the similarity-based group activated a diverse set of regions including bilateral occipital gyrus, left inferior parietal lobes, and bilateral middle frontal gyrus (Figure 9a). Rule-based generalizers recruited the bilateral occipital lobes, the left superior parietal lobes and the left inferior and middle frontal gyrus (Figure 9b). A conjunction analysis indicated common activation between the similarity- and rule-based responders in the bilateral occipital lobes, the left superior parietal lobes, and left postcentral gyrus (see Figure 9c; Table 3). This pattern is, broadly speaking, similar to what we observed in the training phase with the same stimuli.

No differences emerged between the rule and similarity groups in either whole-brain or in our exploratory ROI analyses. This is perhaps not that surprising as these analyses are considerably less sensitive than the analogous critical trials analyses given that they do not directly measure generalization and one cannot determine at the individual trial level whether a response is rule or similarity consistent. As a further exploratory analysis, we used the WFU Pickatlas (Maldjian et al., 2003) to construct a mask containing all regions identified in the Similarity – Rule analysis of the critical items previously reported (Figure 7b), and identified whether there was any activation in these areas for the familiar items at thresholds of $p<.005$ (uncorrected) and 10 contiguous voxels. This analysis revealed activation in the right inferior parietal lobes (peak voxel x = 59, y = -37, z = 31, cluster size: 37

voxels). We also conducted the same type of analysis for the right middle frontal gyrus region (Figure 7a) that was more activated in the rule than the similarity condition for the critical generalization trials but did not detect any activation here.

Participants with a mixture of rule- and similarity-consistent responses

14 participants met the learning criterion for training but showed no clear pattern of similarity- or rule-based responding for the critical generalization trials and were consequently excluded from the analyses above. Nevertheless, given that these participants had a mixture of both types of strategy (rule-consistent, M = .501; similarity consistent, M = .499), they provide an opportunity to look at the neural correlates of rule- and similarity-based responding within-subjects.

Similarity-consistent responses evoked activation in the bilateral anterior cingulate/medial frontal gyrus, the left superior parietal lobe, and the bilateral occipital lobes (Figure 10a). Rule-based responding also activated bilateral occipital lobes (see Figure 10b). We did not, however, detect any significant activation elsewhere although exploratory analyses with more liberal thresholds ($p<.001$ and 25 contiguous voxels) revealed a similar pattern of activation to what we observed with the similarity responders. We suspect the reduced activation here compared to the corresponding analysis for the consistent rule-based sorters reflects the lower number of participants and trials in the current analysis. We found no indication of any differences between strategies in either whole-brain or exploratory ROI analyses.

Discussion

The present study used a negative and positive patterning design originally developed by Shanks and Darby (1998; see also Wills et al., 2011) to compare the brain

activation of rule and similarity generalization. Participants were divided into either rule-based or similarity-based generalizers according to their responses during the critical items at test. Importantly, participants in both groups were highly consistent and did not differ in their ability to follow their preferred strategy. There was extensive overlap of activation between the rule- and similarity-based groups but at the same time there were regions that were differentially activated by the two strategies. We discuss the most notable aspects of our findings below.

In the training phase, there was extensive overlap between groups including diverse regions of the prefrontal cortex, the parietal lobes and the occipital lobes. Differences between strategies, in contrast, were more restricted – no regions were more active in the similarity group than the rule group, although the posterior cingulate/precuneus and the anterior cingulate/ medial prefontal cortex (in the second half of training) were more active in the rule group than the similarity group. The greater posterior cingulate/precuneus activation is perhaps somewhat surprising although these regions have previously been implicated in rule-based category learning (Milton & Pothos, 2011). The anterior cingulate activation is in line with the key role this region is thought to play in rule selection in COVIS's rule-based system (Ashby et al., 1998).

In the test phase, there was again considerable common overlap in the regions activated by the rule and similarity groups for both the critical generalization trials and the familiar trials. As before, areas activated included regions of the prefrontal cortex (including the middle frontal gyrus), bilateral parietal lobes, and bilateral occipital lobes. These regions have all previously been implicated in categorization tasks (e.g., Carpenter et al., in press, Milton et al., 2009). For example, the bilateral parietal lobes have been heavily implicated in both explicit and implicit classification

of dot patterns (Aizenstein et al., 2000) and stimulus generalization (Seger, Braunlich, Wehe, & Liu, 2015). Furthermore, our results add to the growing body of evidence that diverse regions of the prefrontal cortex are involved in categorization (e.g., Grossman et al., 2002; Koenig et al., 2005; Milton & Pothos, 2011; Seger & Cincotta, 2002). The common activation shared by the rule and similarity groups across both the category learning and generalization components of this task is consistent with the idea that both strategies share a number of common, inter-related processes, such as stimulus processing, response selection, stimulus-response mappings, feedback processing (whether it is an external signal as in training or more internally generated as is likely in test), uncertainty and attentional and working memory demands to name just a few likely candidates.

While no differences were identified between generalization strategies in whole-brain analyses, exploratory ROI analyses provided evidence for dissociable activation between the rule and similarity groups. As predicted, the rule-based generalizers activated the right middle frontal cortex to a greater extent than the similarity-based generalizers. This was in keeping with the trend from the training phase for there to be greater activation in the right prefrontal cortex in the rule group than the similarity group. In contrast, the similarity-based generalizers preferentially recruited the anterior medial prefrontal lobe and the right inferior parietal lobes.

The greater right middle frontal gyrus activation in rule-based generalizers than similarity-based generalizers is consistent with a broad range of previous work indicating that the middle frontal lobes is a critical site for rule-based categorization (e.g., Grossman et al., 2002; Milton & Pothos, 2011; Patalano et al., 2001; Seger & Cincotta, 2002, 2005; Tracy et al., 2003) and, more generally, its role is well-established in working memory processing (Owen, 2000). This pattern of findings is,

therefore, in line with the idea that the "opposites" rule is the result of deliberative, rule-based processes, rather than being driven by a non-rule based, associatively mediated system (Maes et al., 2015).

Turning to the regions preferentially linked to similarity responding, the right inferior parietal lobes were identified by Grossman et al. (2002) as being linked to similarity judgments and they suggested that it had a role in overall feature configuration processing. Alternatively, it could reflect recollection-based memory processes which are often observed in this region (e.g., Milton et al., 2011; Wheeler & Buckner, 2004). This would be consistent with the idea that similarity processing places particular demands on the retrieval of past, related, instances. We found that the anterior medial prefrontal lobe was activated more for the similarity than the rule group, in a strikingly similar location to that observed by Koenig et al. (2005) in their analogous comparison. Our explanation for this result is similar to Koenig et al.'s - the activation in this region may reflect the greater dependence on retrieving specific exemplars from long-term memory in the similarity-based condition rather than having generalization supported by an abstract rule.

Of course, there is always a danger in linking specific brain regions to a particular function and in knowing whether the regions identified are essential for the strategy used. For example, the fact that both the similarity and rule groups activated different parts of the prefrontal cortex more than the other strategy appears to challenge any clear-cut narrative that rule generalization requires higher order, deliberative processes while similarity generalization requires more automatic, non-deliberative processes. One intriguing future approach may be to further explore the regions where differences emerged (i.e., the right middle frontal gyrus for the rule generalizers and the anterior medial frontal lobe and right inferior parietal lobes for

the similarity generalizers) using transcranial magnetic stimulation (TMS). For example, according to our results, one might expect that stimulating the right middle frontal gyrus would disrupt rule generalization but leave similarity generalization intact, while stimulating the anterior prefrontal lobe (or right inferior parietal lobes) might impair similarity generalization but not rule generalization. Furthermore, if the greater activation in the anterior medial prefrontal cortex reflects a greater reliance on retrieving past instances in similarity generalization than rule generalization, then one might expect that stimulating this region would disrupt performance in a task using similar stimuli which more directly assesses memory capacity.

Regardless of the precise role that these regions may play, using the Shanks-Darby procedure to examine the neural correlates of rule and similarity generalization makes a valuable new contribution to the area because it overcomes some problems which mar the paradigms used in previous related imaging studies. In particular, many studies in this area confound rule-based and similarity-based learning with single versus multidimensional learning (e.g., Milton & Pothos, 2011; Nomura et al., 2007; Tracy et al., 2003). Furthermore, all previous studies investigating this issue index rule-based learning through behavior that can also be produced by a simple associative mechanism that incorporates some process of selective attention (e.g., ALCOVE, Kruschke, 1992). In contrast, for the Shanks-Darby procedure there is no simple associative model that can explain both the similarity and rule-based generalization observed and the same number of dimensions are relevant in both rule- and similarity-based learning.

There are other notable differences between the Shanks-Darby procedure and the other tasks previously used (e.g., Grossman et al., 2002; Koenig et al., 2005; Nomura et al., 2007; Tracy et al., 2003) which may have impacted the results

obtained. For instance, the Shanks-Darby task enables one to partition items into generalization trials (items not encountered during training) and familiar trials (which had previously been learned during training). In contrast, other studies have either used category learning (e.g., Carpenter et al., in press; Nomura et al., 2007) or category decision making (e.g., Grossman et al., 2002; Milton et al., 2009) procedures where there is no generalization phase, or used imaging analyses which combine old training items with generalization trials (e.g., Koenig et al., 2005). Given that in the test phase, we only observed evidence for differences in the generalization trials and not in the familiar trials this could be an important distinction.

A second important difference is that previous studies typically use multi-dimensional stimuli which possess either binary (e.g., Milton et al., 2009; Tracy et al., 2003) or continuous values (e.g., Carpenter et al., in press; Nomura et al., 2007) on a particular dimension. In contrast, our stimuli have discrete components (e.g., A, B, AB etc) that are either present or absent. This distinction has not previously been systematically investigated but could potentially have an important impact on the pattern of results obtained. Clearly, in future it would be of value to further explore rule and similarity learning under a more diverse range of conditions in order to build up a broader understanding of how the two types of strategies relate to each other.

While the Shanks-Darby procedure appears to have a number of strengths, one potential limitation is that although the rule and similarity groups were well-matched in their consistency of applying their preferred strategy for the critical generalization trials, the rule group significantly outperformed the similarity group for the familiar training trials (this is the same pattern which Shanks & Darby, 1998, found). It is worth noting, though, that the familiar trials were analyzed separately from the critical generalization trials where the neural differences emerged, which should attenuate any

influence this difference would have on our results. Furthermore, differences between groups for the familiar items (where only correct responses were considered) were extremely restricted which suggests that performance difference between groups had little impact on the pattern of activation. Nevertheless, in future work with the Shanks-Darby procedure it would be desirable to have the groups better matched on performance for the familiar trials. One way of doing this could be to introduce a learning criterion during the training phase, which has been shown to better equate groups on the familiar test items in this procedure (Wills et al., 2011).

Another notable aspect of our results is that the key comparison between rule and similarity generalization was between-subjects. While we looked at this from a within-subjects perspective as well by considering those participants who displayed a mixture of strategies, these analyses were not particularly revealing. We suspect that this was due to these participants randomly responding on the critical trials and/or that there were insufficient trials of each type to reliably detect any differences. In future, one could potentially further increase the number of critical generalization trials in the procedure, and/or use a combination of more effective instructions than we used with training outside the scanner to ensure that participants produce a good mixture of rule and similarity generalization responses.

Another limitation of this study is that while the differences between generalization strategies were consistent with our a priori predictions and in line with past related work, it must be acknowledged that these differences were not identifiable in whole-brain analyses and could only be detected in more exploratory, post-hoc analyses. Clearly, it would be valuable for future work to try and replicate our basic pattern of results.

Nevertheless, our finding of differences in activation between rule and similarity generalization (albeit only in exploratory analyses) is in line with recent behavioral and comparative studies that have been conducted with the Shanks-Darby procedure. Specifically, Wills et al. (2011) found that participants who completed the training session under a concurrent load went on to produce significantly less rule-based generalization than participants who undertook training under no load. Wills et al. suggested that this is consistent with the idea that discovering the "opposites" rule requires considerable working memory capacity and if this is not available that participants will fail to transition from using a similarity approach to a rule-based approach. Related to this, Maes et al. (2015) found that while humans (under no concurrent load) could readily discover the opposites rule, both pigeons and rats were unable to do so and relied on similarity generalization. This is consistent with the idea that pigeons and rats may be forced to rely on the similarity system while humans also have access to a rule-based system. Our findings complement these two recent behavioral studies by identifying specific neural correlates that are associated with rule and similarity-based generalization.

However, while our results make a novel and valuable contribution to the area, our findings stop some way short of providing clear evidence for qualitatively separable generalization systems by any reasonable definition. For instance, neural differences in themselves should not be taken as evidence for separable systems given that past work has shown that items within the same category can provoke differential activation (e.g., Davis & Poldrack, 2014; DeGutis & D'Esposito, 2009; Grinband et al., 2006).What may be more compelling evidence for separable learning systems would be large differences in activation in regions that are not also activated by the other strategy. This does not appear to be the case in the present study where the

differences in generalization were restricted and located close to regions that were activated in training and test by both strategies. Furthermore, the commonalities in activation between strategies clearly outweigh the differences.

One potentially fruitful way of viewing the current data, then, is to consider category learning and stimulus generalization as cognitively complex processes comprising a number of sub-components (e.g., stimulus processing, hypothesis testing, decision making, feedback processing etc) many of which are likely to be shared by rule and similarity strategies. Furthermore, one strategy may place more of an emphasis on one of these sub-processes than the other strategy does. For example, the "opposites" rule needed for rule generalization appears likely to place particular demands on working memory capacity and rule formation that do not appear to be needed to the same extent for similarity generalization and this may require increased activation of the right middle frontal gyrus. Conversely, a similarity strategy may, for example, impose higher memory demands than the rule condition (where there may be more of an emphasis on abstract rules) which could lead to greater engagement of the anterior medial prefrontal lobes. Of course, our hypotheses as to what particular role these brain regions play may not be correct but as others have recommended (e.g., Davis et al., 2012) trying to link brain regions to specific functions of the learning process may be at least as profitable an approach as focusing on the more general and contentious question of whether data is more in line with single or multiple system accounts. We suggest that further examination of the Shanks-Darby procedure, with its notable strengths, could play an important role in further illuminating both these important questions.

References

Aizenstein, H.S., MacDonald, A.W., Stenger, V.A., Nebes, R.D., Larson, J.K., Ursu, S., & Carter, C.S. (2000). Complementary category learning systems using event-related functional MRI. *Journal of Cognitive Neuroscience, 12*, 977-987.

Allen, S.W. & Brooks, L.R. Specializing the operation of an explicit rule. (1991). *Journal of Experimental Psychology: General, 120*, 3-19.

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

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

Baguley, T. & Kaye, D. (2010). Book review: Understanding psychology as a science: An introduction to scientific and statistical inference. *British Journal of Mathematical and Statistical Psychology, 63*, 695-698.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization.* Hillsdale, N.J.: Erlbaum.

Carpenter, K.L., Wills, A.J., Benattayallah, A., & Milton, F. (in press). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*. DOI: 10.1002/hbm.23259

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

Davis, T., & Poldrack, R.A. (2014). Quantifying the internal structure of categories using a neural typicality measure. *Cerebral Cortex, 24*, 1720-1757.

DeGutis, J. & D'Esposito, M. (2009). Network changes in the transition from initial learning to well-practiced visual categorization. *Frontiers in Human Neuroscience, 3*(44), 1-13.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives in Psychological Science, 6*, 274-290.

Edmunds, C.E.R., Milton, F., & Wills, A.J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category learning. *Quarterly Journal of Experimental Psychology, 68*, 1203-1222.

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

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *Neuroimage, 43*, 509-520.

Grossman, M., Smith, E.E., Koenig, P., Glosser, G., DeVita, C., Moore, P. & McMillan, C. (2002). The neural basis for categorization in semantic memory. *Neuroimage, 17*, 1549–1561.

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

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning &Verbal Behavior 23*, 734-759.

Kiebel, S.J., Poline, J.B., Friston, K.J., Homes, A.P., & Worsley, K.J. (1999). Robust smoothness estimation in statistical parametric mapsusing standardized residuals from the general linear model. Neuroimage, 10, 756-766.

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

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Maes, E., De Filippo, G., Inkster, A.B., Lea, S.E.G. De Houwer, J., D'Hooge, R., Beckers, T., & Wills, A.J. (2015). Feature-versus rule-based generalization in rats, pigeons and humans. *Animal Cognition, 18*, 1267-1284.

Maldjian, J.A., Laurienti, P.J., Burdette, J.B., Kraft, R.A. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage, 19,* 1233–1239.

Milton, F., & Wills, A. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30*, 407-415.

Milton, F., Longmore, C.A., & Wills, A.J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 676-692.

Milton, F., & Pothos, E.M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience, 34,* 1326-1336.

Milton, F., Wills, A.J., & Hodgson, T.L. (2009). The neural basis of overall similarity and single- dimension sorting. *Neuroimage 46,* 319–326.

Milton, F., Muhlert, N., Butler, C.R., Benattayallah, A., & Zeman, A.Z.J. 2011. The neural correlates of everyday recognition memory. *Brain and Cognition, 76*, 369-381.

Newell, B.R., Dunn, J.C., & Kalish, M. (2011). 6 Systems of Category Learning: Fact or Fantasy? In B.H. Ross (Ed). *The Psychology of Learning & Motivation, 54*, 167-215.

Newell, B.R., Moore, C.P., Wills, A.J., & Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves implicit perceptual categorization: a comment on Filoteo, Lauritzen, and Maddox (2010). *Psychological Science*, *24*, 386-389.

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

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

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

Nosofsky, R.M., & Johansen, M.K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7*, 375-402.

Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review, 9*, 169-174.

Owen, A.M. (2000). The role of the lateral frontal cortex in mnemonic processing: The contribution of functional imaging. *Experimental Brain Research, 133*, 33-43.

Patalano, A.L., Smith, E.E., Jonides, J., & Koeppe, R.A. (2001). PET evidence for multiple strategies of categorization. *Cognitive, Affective & Behavioral Neuroscience, 1*, 360-370.

R, Core Team (2015). R: A Language and Environment for Statistical Computing. http://www.R-project.org/, R Foundation for Statistical Computing Vienna, Austria.

Rips, L.J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.) *Similarity and analogical reasoning* (pp.21-59). Cambridge: Cambridge University Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

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

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

Seger, C.A. & Cincotta, C.M. (2002). Striatal activity in concept learning. *Cognitive Affective & Behavioral Neuroscience, 2*, 149–161.

Seger, C.A. & Cincotta, C.M. (2005). The roles of the caudate nucleus in human classification learning. *Journal of Neuroscience, 25*, 2941–2951.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24*, 405- 415.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65*, 167-196.

Stanton, R.D., Nosofsky, R.M. (2013). Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 1174-1191.

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

Tracy, J.I., 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 and Cognition*, 51, 276–286.

Ward, T.B. & Scott, J. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, *15*, 42-54.

Wheeler, M.E., & Buckner, R.L. (2004). Functional-anatomic correlates of remembering and knowing. *Neuroimage, 21*, 1337-1349.

Wilkinson, D. T., Halligan, P. W., Henson, R. N. A., and Dolan, R. J. (2002). The effects of interdistracter similarity on search processes in the superior parietal cortex. *Neuroimage 15*, 611–619.

Wills, A.J., Graham, S., Koh, Z., Mclaren, I.P.L., & Rolland, M.D. (2011). Effects of concurrent feature- and rule-based generalization in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 308-316.

Wills, A.J., Inkster, A.B., & Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. *Cognitive Psychology, 80*, 1-33.

Wills, A.J., Milton, F., Longmore, C.A., Hester, S., Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *Quarterly Journal of Experimental Psychology, 66*, 299-318.

Footnotes

[1] The raw imaging and behavioral data is available for interested readers at: http://www.willslab.co.uk/exe10/index.html.

[2] By convention, a Bayes factor (BF) of over three is interpreted as providing substantial evidence for the experimental hypothesis (Jeffreys, 1961), while a BF below a third provides substantial evidence for the null (Dienes, 2011). BF analysis requires an estimate of the mean expected difference under the experimental hypothesis; we estimated this from the observed difference for the familiar test items. Following Dienes (2011), the expected difference was modeled as a two-tailed normal distribution with a standard deviation equal to half the mean. Calculations were run using a custom script (Baguley & Kaye, 2010) within R (R Core team, 2015).

Acknowledgments

Correspondence should be addressed to Fraser Milton, Discipline of Psychology, University of Exeter, Exeter, EX4 4QG. E-mail: f.n.milton@ex.ac.uk. We thank the Experimental Psychology Society for their support and two anonymous reviewers for their insightful and constructive comments.

Table 1

Regions Commonly Activated by Rule-Based and Similarity-Based Generalization for in the Training Phase.

			Talairach Coordinates			
Region	*Cluster size*	*BA*	*x*	*y*	*z*	*z-score*
Left anterior cingulate	1176	32	-6	18	40	6.59
Right anterior cingulate		32	6	16	40	5.98
Left medial frontal gyrus		32	-8	12	45	5.89
Left middle frontal gyrus	5721	9	-50	7	29	6.40
Left precuneus		19	-28	-67	29	6.36
Left inferior parietal lobe		40	-32	-50	43	6.34
Right inferior occipital gyrus	3147	18	38	-82	-4	5.83
Right occipital lobe		19	28	-74	-5	5.71
Right middle occipital gyrus		18	36	-84	2	5.66
Left insula	185	13	-30	20	1	5.65
Left inferior frontal gyrus		47	-30	25	-5	4.90
Right inferior frontal gyrus	196	45	32	24	4	5.34
Right superior parietal lobe	406	7	32	-52	45	4.68
Right precuneus		7	18	-62	49	4.19
Right superior parietal lobe		7	32	-58	40	4.19
Right precentral gyrus	354	4	48	-11	48	4.66
Right precentral gyrus		6	34	-14	62	4.18
Right precentral gyrus		6	40	-7	52	3.87

Note. BA = brodmann's area. All activations significant at $p<.001$. Indented rows indicate voxels in the same cluster as the non-indented row above them.

Table 2

Regions Commonly Activated by Rule-Based and Similarity-Based Generalization for in the Critical Generalization Trials

			Talairach Coordinates			
Region	*Cluster size*	*BA*	*x*	*y*	*z*	*z-score*
Right superior parietal lobe	566	7	32	-58	51	5.40
Right precuneus		19	30	-68	29	4.77
Right inferior parietal lobe		40	34	-50	41	4.36
Left superior parietal lobe	1040	7	-26	-62	44	5.09
Left precuneus		7	-26	-67	27	5.05
Left parietal lobe		39	-28	-62	36	4.97
Left middle occipital gyrus	556	18	-26	-91	10	4.82
Left middle occipital gyrus		18	-26	-84	-4	4.74
Left inferior occipital gyrus		19	-38	-76	-5	4.23
Right middle occipital gyrus	425	18	24	-89	10	4.44
Right middle occipital gyrus		18	16	-94	14	4.21
Right occipital lobe		17	22	-88	-2	3.97
Left medial frontal gyrus	329	6	-2	16	45	4.40
Right medial frontal gyrus		6	8	16	42	3.94
Left middle frontal gyrus	104	9	-50	4	33	3.63
Left precentral gyrus		6	-38	0	33	3.46

Note. BA = brodmann's area. All activations significant at $p<.001$. Indented rows indicate voxels in the same cluster as the non-indented row above them.

Table 3

Regions Commonly Activated by the Rule and Similarity Groups During Test for the Familiar Items.

			Talairach Coordinates			
Region	*Cluster size*	*BA*	*x*	*y*	*z*	*z-score*
Right superior parietal lobe	154	7	30	-49	39	5.37
Right parietal lobe		39	30	-56	38	4.70
Left middle occipital gyrus	565	18	-36	-86	-2	5.17
Left middle occipital gyrus		18	-28	-80	-6	5.08
Left cuneus		17	-14	-91	8	5.01
Left superior parietal lobe	858	7	-30	-56	40	5.07
Left precuneus		19	-28	-68	38	4.88
Left superior parietal lobe		7	-30	-62	49	4.81
Right middle occipital gyrus	564	18	14	-91	14	4.47
Right middle occipital gyrus		18	38	-82	1	4.44
Right middle occipital gyrus		18	30	-87	4	4.33
Left anterior cingulate	244	32	-8	17	38	4.45
Right superior frontal gyrus		8	2	18	47	3.86
Left superior frontal gyrus		6	-6	8	49	3.54

Note. BA = brodmann's area. All activations significant at $p<.001$. Indented rows indicate voxels in the same cluster as the non-indented row above them.

Figure Captions

Figure 1: The training and test trial types in the Shanks and Darby (1998, Experiment 2) allergy prediction task; letters indicate foods eaten by a hypothetical patient Mr. X, + = patient develops an allergic reaction; - = patient does not develop an allergy reaction; ? = no feedback given.

Figure 2: a) Accuracy of rule-based and similarity-based responders for familiar items during the test phase; b) Proportion of critical trials across blocks which are consistent with the strategy participants were assigned to.

Figure 3: a) Mean probability of predicting an allergic reaction for the critical compound stimuli; b) Mean probability of predicting an allergic reaction for the critical element stimuli. Also shown are difference-adjusted 95% confidence intervals for the between-subjects effects (Baguley, 2012).

Figure 4: a) Brain regions activated across all trials and all participants during training; b) Regions activated by correct responses by the rule group during training; c) Regions activated by correct responses in the similarity group during training; d) Common brain regions activated by correct responses in the similarity and rule groups The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Figure 5: a) Brain regions more activated by the rule group than the similarity group across all blocks of training; b) Brain regions more activated by rule-based responders

than similarity responders across the second half of training. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

Figure 6: a) Brain regions significantly activated by similarity responders during the critical trials; b) Brain regions significantly activated by rule-based responders during the critical trials; c) Common brain regions activated by similarity and rule-based responders during the critical trials. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

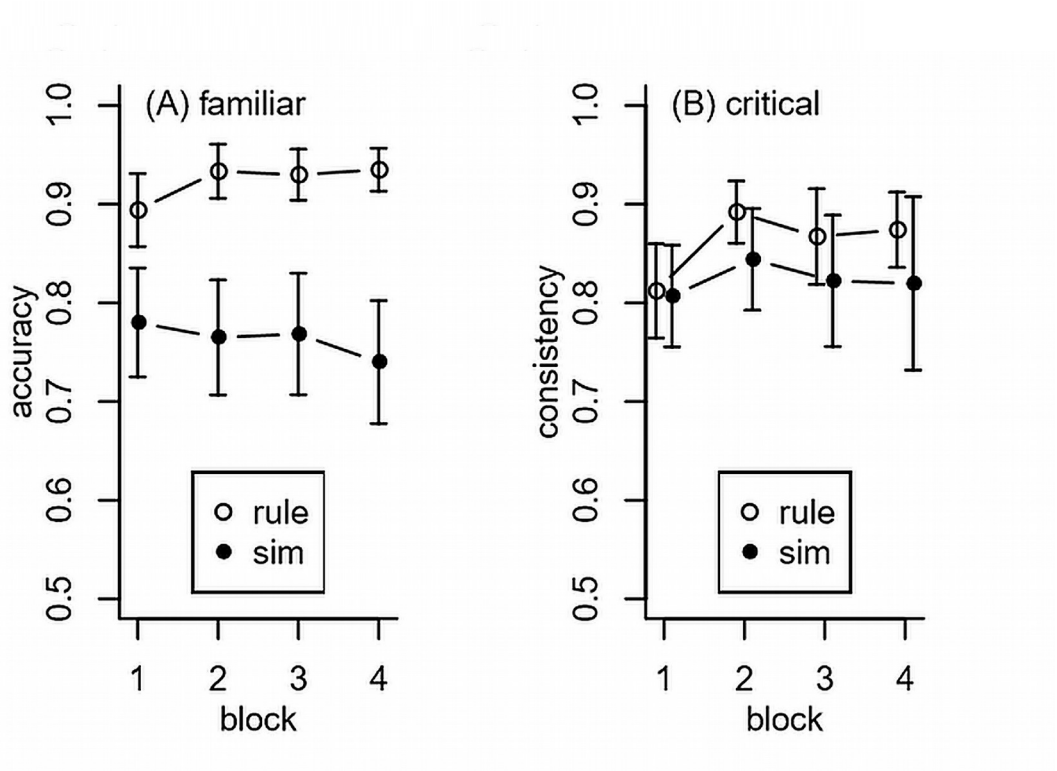
Figure 7: a) Brain regions more activated by the rule-based responders than similarity responders for the critical trials; b) Brain regions more activated by similarity responders than rule-based responders for the critical trials. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

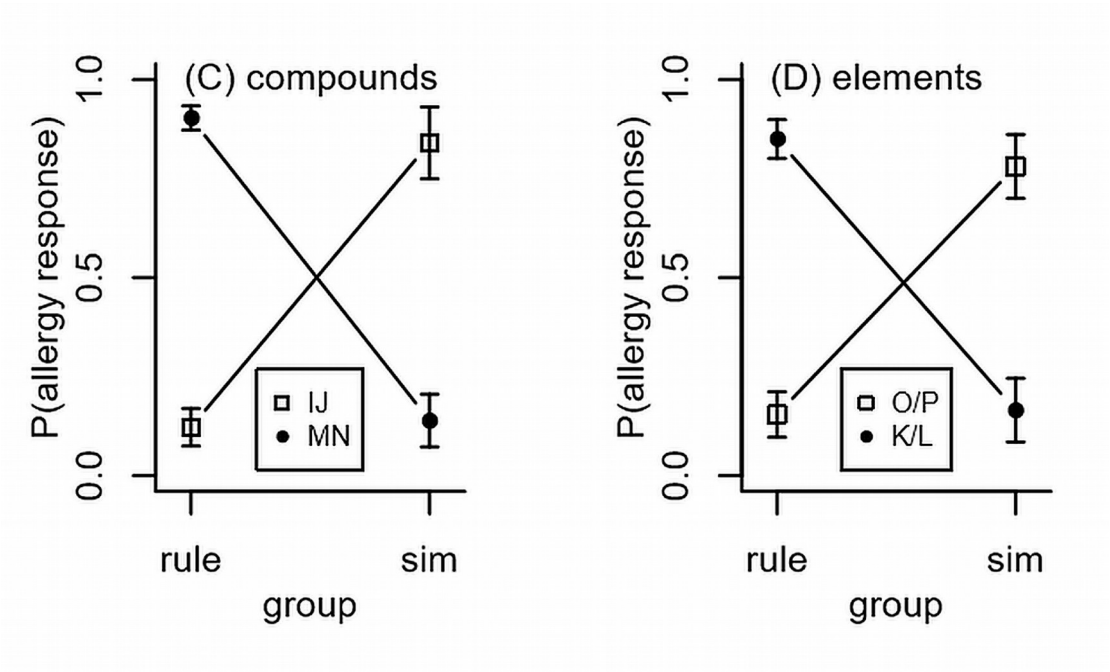
Figure 8: a) Regions more activated by compound items than element items during the critical generalization trials for the rule-based group; b) Regions more activated by element items than compound items during the critical generalization trials for the similarity group.

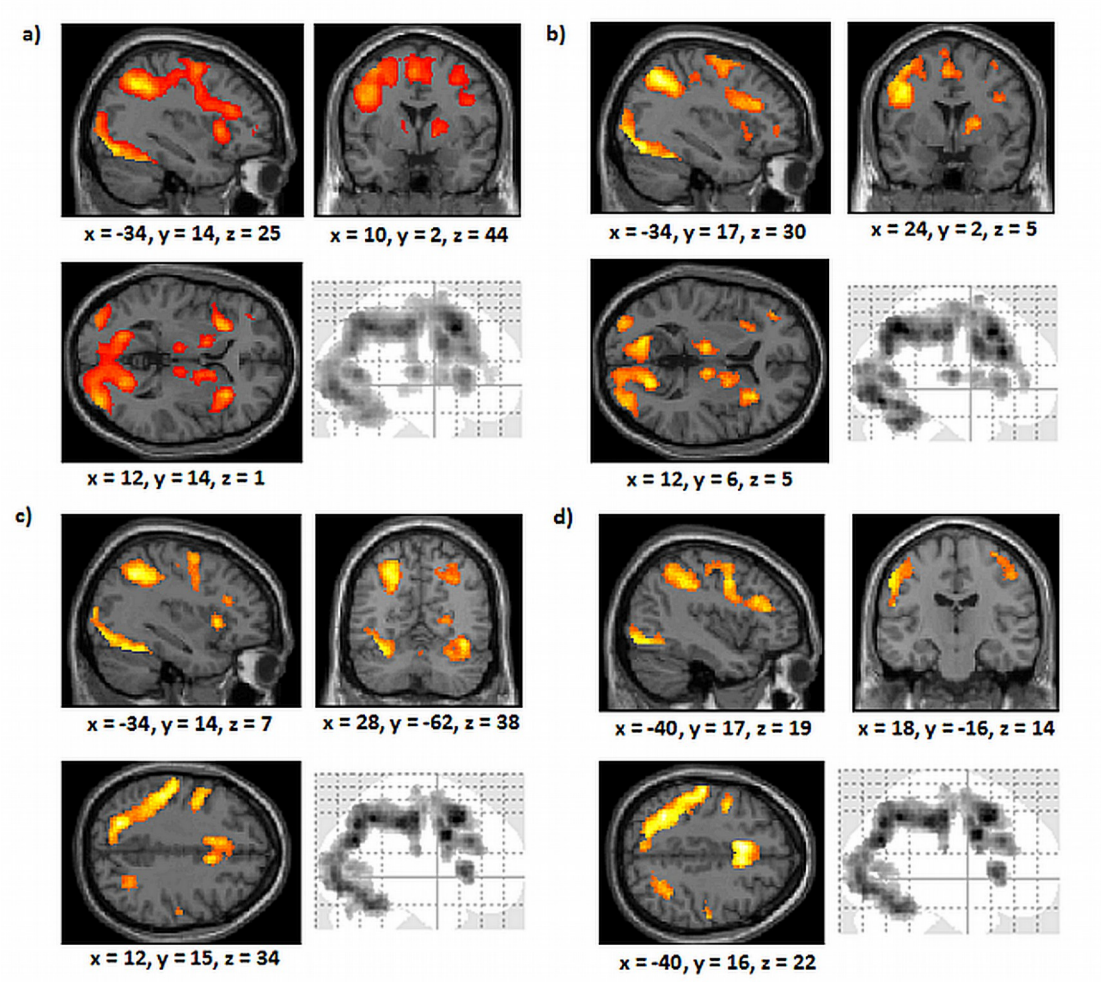
Figure 9: a) Brain regions engaged by the similarity responders for the familiar items during test; b) Brain regions engaged by the rule-based responders for the familiar items during test; c) Brain regions commonly engaged by the similarity and rule-based responders for the familiar items during test. The coordinates indicate the origin for the image displayed. Lighter colors indicate higher z-scores.

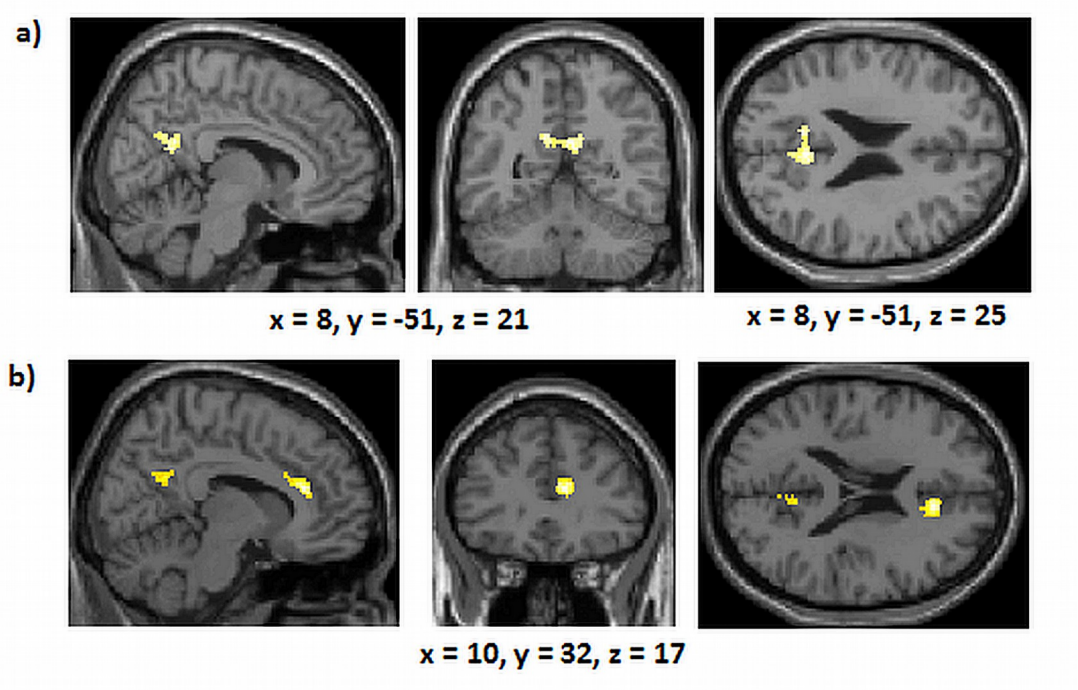
Figure 10: a) Regions activated by similarity-consistent responses for the group who displayed a mixture of both strategies; b) Regions activated by rule-consistent responses for the group who displayed a mixture of both strategies.

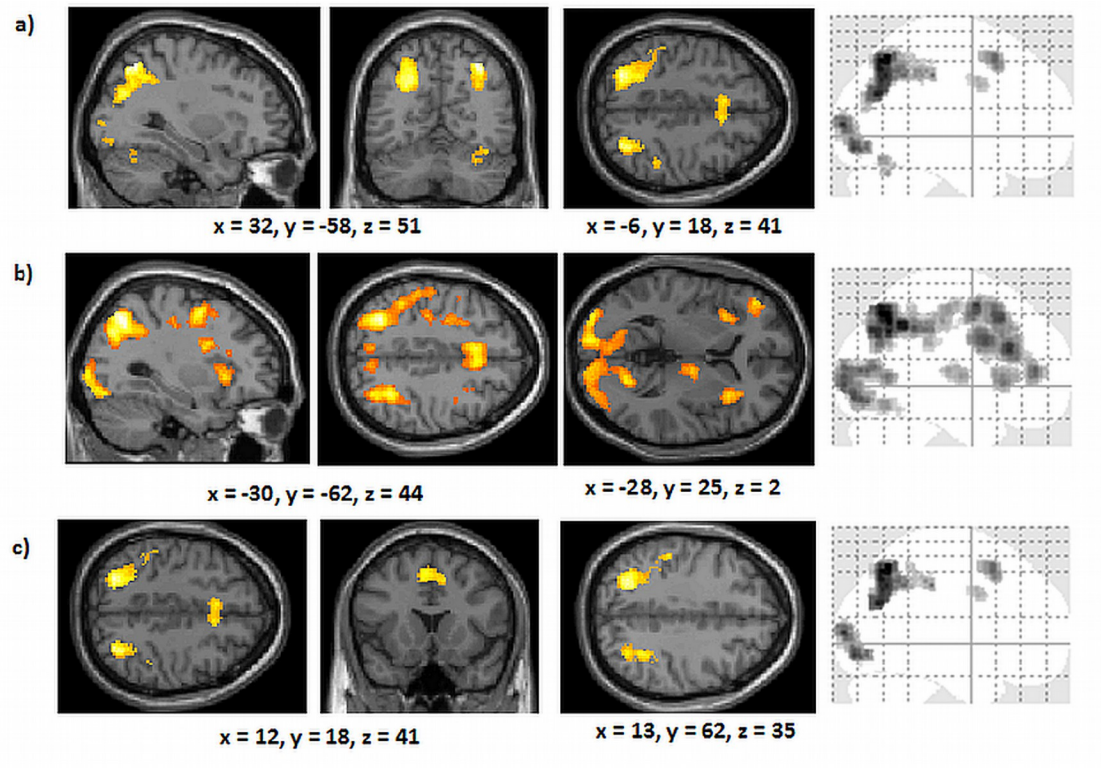
Training			Test		
A+	B+	AB-	A?	B?	AB?
C-	D-	CD+	C?	D?	CD?
E+	F+	EF-	E?	F?	EF?
G-	H-	GH+	G?	H?	GH?
I+	J+		I?	J?	**IJ?**
		KL-	**K?**	**L?**	KL?
M-	N-		M?	N?	**MN?**
		OP+	**O?**	**P?**	OP?

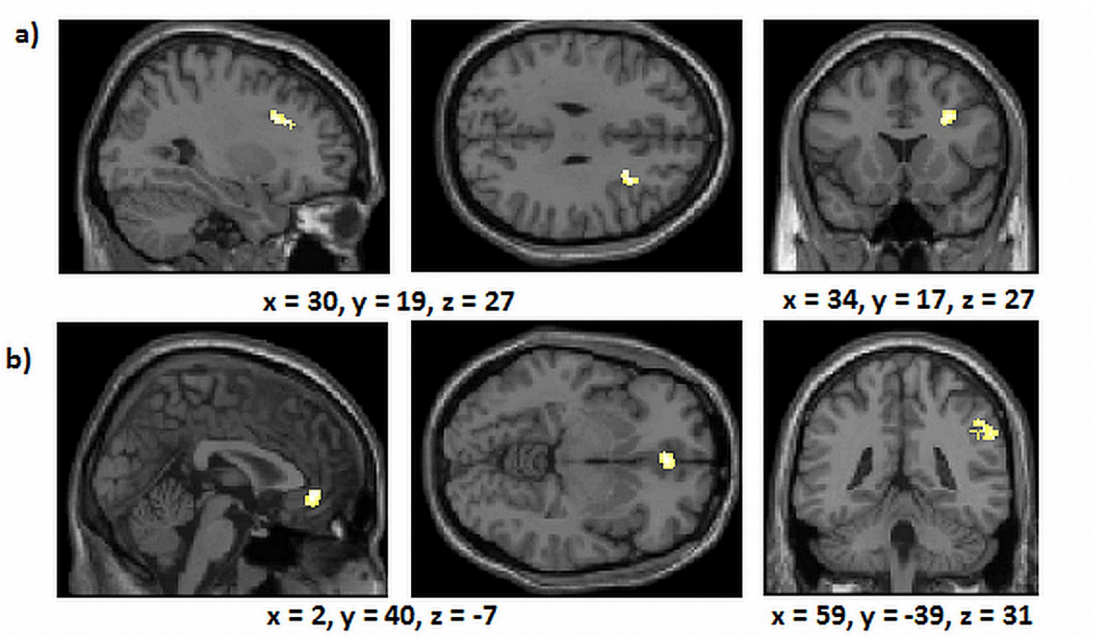

(A) familiar
accuracy
0.5
0.6
0.7
0.8
0.9
1.0
rule
sim
1
2
3
4
block
(B) critical
consistency
0.5
0.6
0.7
0.8
0.9
1.0
rule
sim
1
2
3
4
block

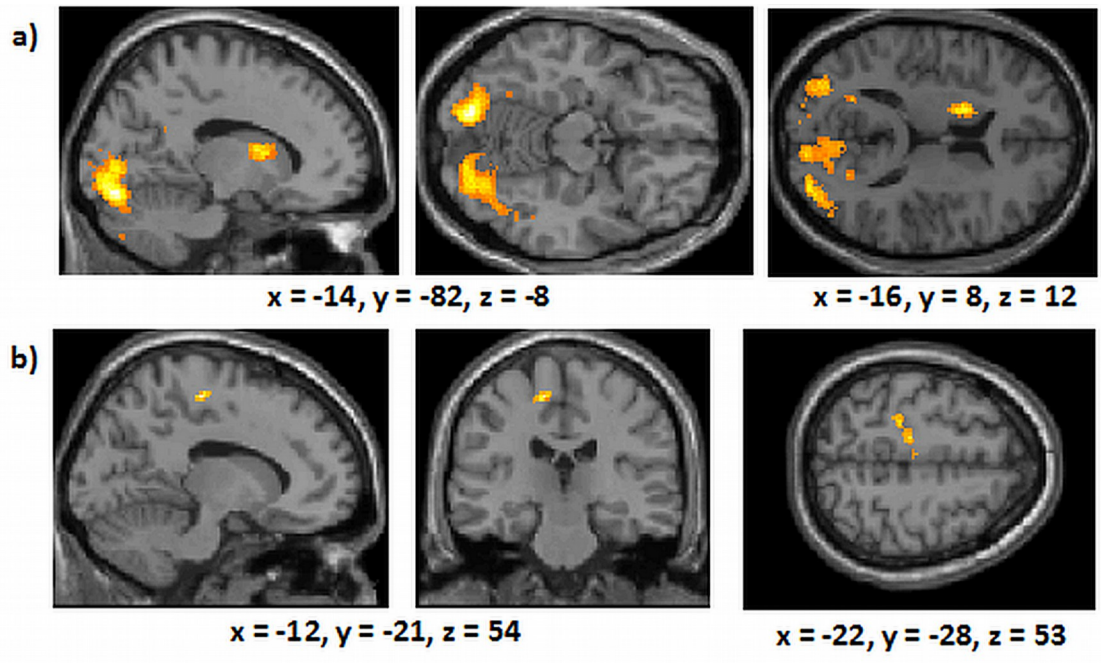

(C) compounds
P(allergy response)
0.0
0.5
1.0
IJ
MN
rule
sim
group
(D) elements
P(allergy response)
0.0
0.5
1.0
O/P
K/L
rule
sim
group

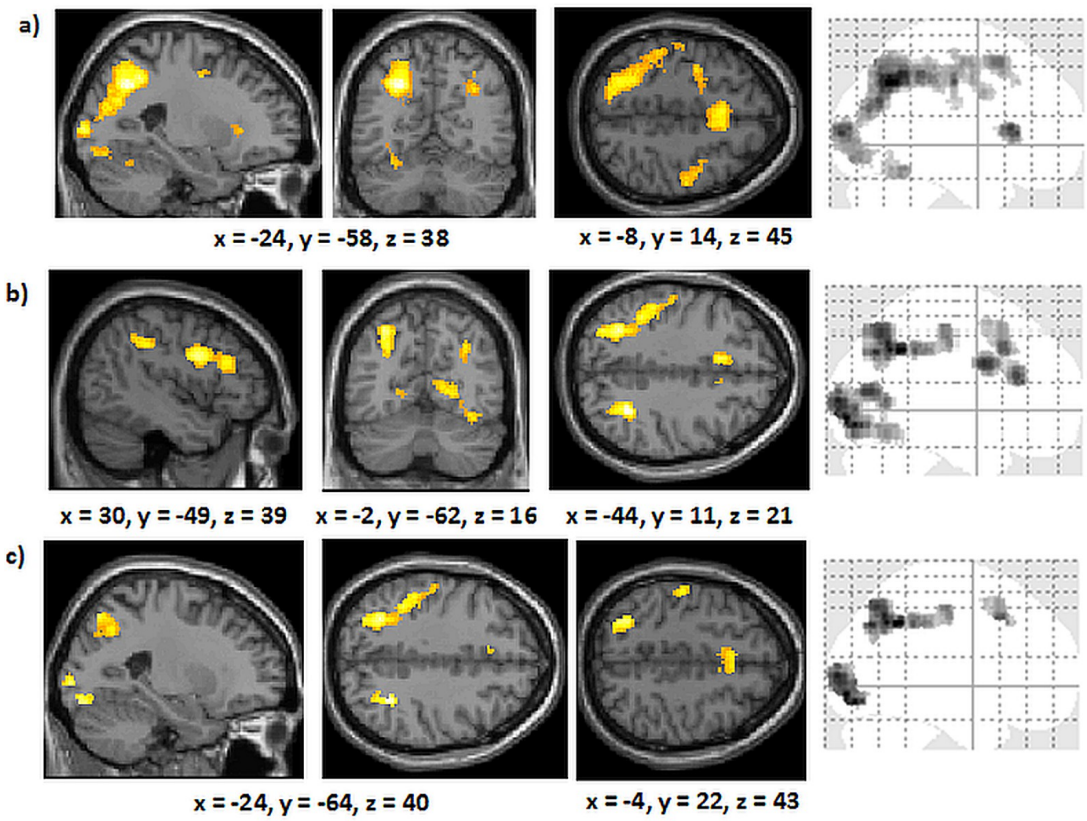

a)
x = -34, y = 14, z = 25
x = 10, y = 2, z = 44
x = 12, y = 14, z = 1
b)
x = -34, y = 17, z = 30
x = 24, y = 2, z = 5
x = 12, y = 6, z = 5
c)
x = -34, y = 14, z = 7
x = 28, y = -62, z = 38
x = 12, y = 15, z = 34
d)
x = -40, y = 17, z = 19
x = 18, y = -16, z = 14
x = -40, y = 16, z = 22

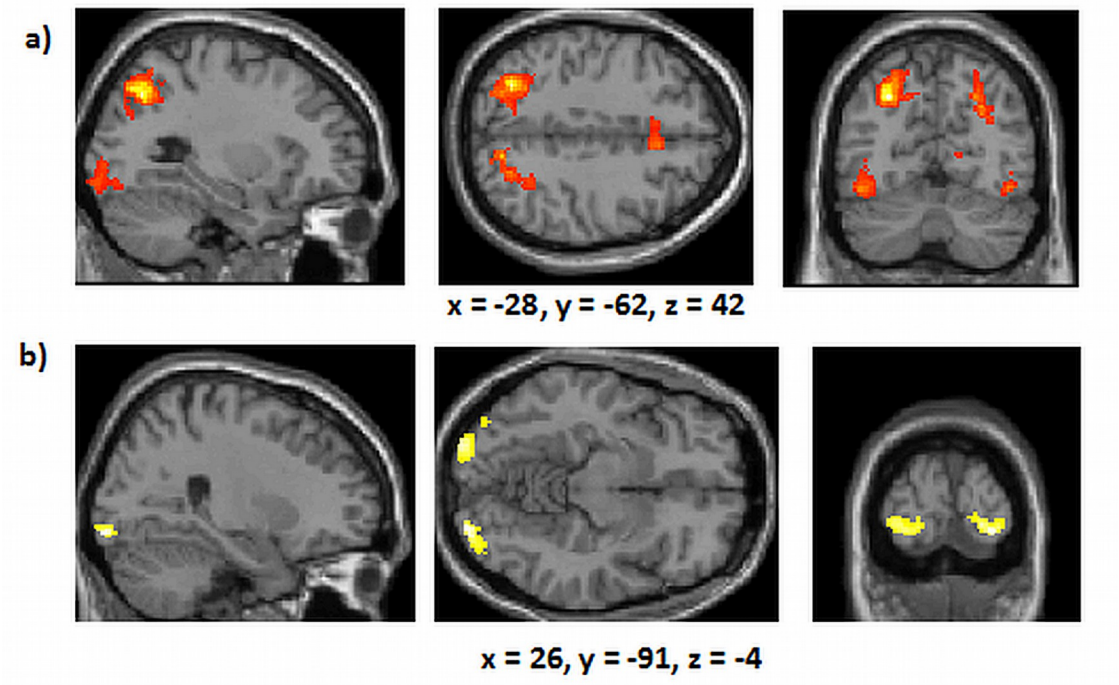

a)
x = 8, y = -51, z = 21
x = 8, y = -51, z = 25
b)
x = 10, y = 32, z = 17


a)
x = 32, y = -58, z = 51
x = -6, y = 18, z = 41
b)
x = -30, y = -62, z = 44
x = -28, y = 25, z = 2
c)
x = 12, y = 18, z = 41
x = 13, y = 62, z = 35


a)
x = 30, y = 19, z = 27
x = 34, y = 17, z = 27
b)
x = 2, y = 40, z = -7
x = 59, y = -39, z = 31


a)
x = -14, y = -82, z = -8
x = -16, y = 8, z = 12
b)
x = -12, y = -21, z = 54
x = -22, y = -28, z = 53


a)
x = -24, y = -58, z = 38
x = -8, y = 14, z = 45
b)
x = 30, y = -49, z = 39
x = -2, y = -62, z = 16
x = -44, y = 11, z = 21
c)
x = -24, y = -64, z = 40
x = -4, y = 22, z = 43


a)
x = -28, y = -62, z = 42
b)
x = 26, y = -91, z = -4