Mathematical models of visual category learning enhance fMRI data analysis

Emi M Nomura (e-nomura@northwestern.edu)
Department of Psychology, 2029 Sheridan Road
Evanston, IL 60201 USA

W Todd Maddox (maddox@psy.utexas.edu)
Department of Psychology, 1 University Station
Austin, TX 78712 USA

Paul J Reber (preber@northwestern.edu)
Department of Psychology, 2029 Sheridan Road
Evanston, IL 60201 USA

Abstract

Models of categorization enable us to test specific hypotheses about psychological processes. Neuroimaging provides us with the tools to visualize the neural correlates of these same processes in human subjects. By combining these techniques, we can begin to make the connection between behavior and neural activity. Here we collected fMRI data in a category learning paradigm where subjects were encouraged toward a particular strategy by the underlying category structure. The application of mathematical models of category learning to this behavioral data enabled us to better organize the fMRI data according to the actual strategy employed.

Keywords: categorization; modeling; neuroimaging

Categorization is a process by which the brain assigns meaning to stimuli. Through experience with these stimuli, we learn to group distinct items that may share similar properties into categories, such as 'cars' or 'birds.' This grouping is critical for rapidly and appropriately selecting behavioral responses. While there are numerous categories that we've already acquired over our lifetime, as adults we possess the ability to acquire new categories. It is this learning process that is under investigation here.

The study of categorization has a long history in cognitive psychology, particularly in the domain of computational and mathematical modeling. A variety of models exist that support different theories about the cognitive operations involved in categorization. These models are traditionally tested by comparison to human behavioral data in an attempt to understand the specific operations that contribute to categorization.

Collecting functional magnetic resonance imaging (fMRI) data while subjects are undergoing category learning allows us to test hypotheses about the involvement of neural processes in theoretical models of categorization. Recently, a number of different neuroimaging studies have investigated the neural correlates of category learning in different types of tasks. In general, the data suggests that there exist multiple neural systems that can support category learning.

One useful approach aimed at dissociating these systems is to present different category structures that preferentially elicit one learning system over another. fMRI can then be used to visualize the specific patterns of functional activity associated with trials where learning took place. However, this approach presupposes that learning the categories only recruits the operation of one system at a time. A key question is whether the systems are independent or competitive.

Here we collected fMRI data from two groups performing visual category learning in two different ways. The underlying category structure of the to-be-learned categories differed between groups so as to encourage one type of strategy over another. Mathematical models of learning were fit to the behavioral data to identify the specific strategy each subject was using. Modeling identified particularly good instances of strategy use, but also demonstrated that some subjects used the sub-optimal strategy regardless of their experimental group. We then analyzed the fMRI data in two different ways. The first grouped subjects according to their experimental group, which assumed that the group assignment was sufficient to elicit the appropriate learning strategy. The second re-grouped the subjects according to their preferred strategy. By looking selectively at subjects who best exhibited each strategy, additional elements of the neural circuits supporting each type of category learning were identified. Our results demonstrate that models of category learning can effectively guide fMRI data analysis. Likewise, we can take what we've learned from imaging and use it to improve upon existing models of categorization.

Models of categorization

In the literature, models designed to capture category learning behavior can be grouped into three general types: exemplar, prototype and decision-boundary models.

Exemplar models of categorization (Medin and Schaffer 1978; Hintzman 1986; Nosofsky 1986) posit that people represent categories by storing individual exemplars of the category in memory. When encountering a novel item, their classification decision depends on the similarity of that item to every stored exemplar. Categorization depends upon memory for these exemplars, which is presumably the same

memory that allows for the recognition of those exemplars (Nosofsky 1991). One major challenge to exemplar theory comes from the performance of amnesic patients on a categorization task. These patients have no memory for individual stimuli, but retain the ability to categorize (Knowlton and Squire 1993), which refutes this single-system view.

Prototype models (Posner and Keele 1968; Reed 1972; Smith, Murray et al. 1997) maintain that a category representation consists of a prototype of the trained exemplars. In this case, a prototype is usually defined as corresponding to the central tendency of the experienced stimuli. Novel stimuli are then categorized according to their similarity to this prototype. This differs from the exemplar models in that one does not need to have seen the prototype of the category to effectively learn, rather this information is accumulated over training.

In decision-bound models (Ashby and Townsend 1986), people make categorical decisions using a decision boundary that divides a multidimensional psychological space into category-response regions. Through experience, the categorizer learns to associate a particular category label with each region, and learning the categories amounts to identifying the decision-boundary that separates the categories. The mathematical models described here are based on the decision-bound theory of categorization.

Decision-bound theory

A number of reports support DBT as an effective description of visual category learning (Ashby and Gott 1988; Ashby and Maddox 1990; Ashby and Maddox 1992; Maddox and Ashby 1993). Typically, the stimuli in these experiments vary on 2 or more dimensions. For example, in the current task the stimuli were sine wave grating that varied in frequency (thickness of lines) and orientation of lines. The two-dimensional perceptual space of the stimuli can be partitioned into 2 (or more) categories by decision boundaries that can be linear or non-linear.

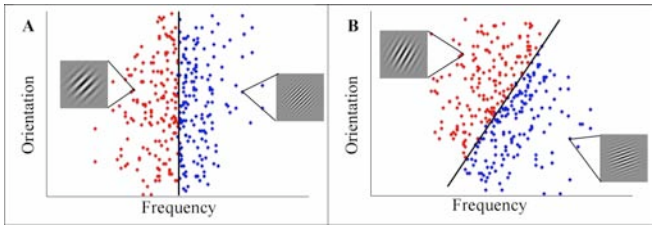

Figure 1: (A) RB and (B) II category learning tasks. Each point represents a distinct Gabor patch (sine-wave) stimulus defined by orientation and frequency (thickness of lines). In both stimulus sets, there are 2 categories (red and blue points). RB categories are defined by a vertical boundary (only frequency is relevant for categorization) whereas II categories are defined by a diagonal boundary (both dimensions are relevant). In both RB and II stimuli there are examples of a stimulus from each category.

A linear boundary that segments the perceptual space along one dimension (e.g., a horizontal or vertical boundary) can be easily described by a verbal rule (Rule-based; RB). In contrast, a decision boundary that does not fall along a cardinal orientation requires the learner to integrate information across the 2 dimensions in order to determine category membership (Information-integration; II). These two category structures (Figure 1), while existing in the same perceptual space, require very distinct types of learning strategies.

COVIS

One multiple system model that provides a specific hypothesis about the neural basis of RB and II categorization is the COVIS model (COmpetition between Verbal and Implicit Systems) proposed by Ashby (Ashby, Alfonso-Reese et al. 1998). In this model, 2 learning systems compete to provide the output response: an explicit, rule-based system dependent upon working memory and attention; and an implicit, procedural learning system. While COVIS itself is not a DBT model, we can use DBT-based models to test the theory. Evidence in favor of COVIS comes from a number of sources, most recently neuroimaging.

Under the COVIS theory, the rule-based system learns through a conscious process of rule generation and testing, cognitive functions normally subserved by the frontal lobes. Neuroimaging of rule-based tasks has shown consistent activity in the prefrontal cortex (PFC), anterior cingulate and head of the caudate (Rao, Bobholz et al. 1997; Lombardi, Andreason et al. 1999; Filoteo, Maddox et al. 2005). The theory maintains that potential rules being tested are stored in working memory and are either discarded or retained according to the feedback. The structures within the medial temporal lobe (MTL) may also support this type of learning (Ashby and Valentin in press) by maintaining the specific rule that distinguishes the categories. This could effectively consist of memorizing a specific stimulus near the decision boundary.

The implicit learning system in COVIS is hypothesized to depend upon the posterior body and tail of the caudate nucleus and its interconnections with posterior visual cortical areas. Evidence supporting the important role of the caudate in this system is based primarily on its neurobiological properties. The spiny neurons in the tail of the caudate receive projections from the visual cortical neurons in TE (inferotemporal cortex) in a many-to-one fashion (Wilson 1995). This massive convergence allows a wide variety of complex information to be compressed to its most basic representation, which is precisely the type of process necessary for categorization.

While neurally inspired, the COVIS model makes several strong predictions about the different behavioral characteristics of RB and II category learning that have been tested empirically. RB category learning was disrupted more than II category learning by the simultaneous performance of either a numerical Stroop task (Waldron and Ashby 2001)

or a sequential memory scanning task (Maddox, Ashby et al. 2004). The interfering tasks both rely on working memory and attention, thus supporting the importance of these processes to the RB system. In contrast, manipulations of the nature and timing of the feedback impact II category learning more than RB category learning (Maddox, Ashby et al. 2003; Maddox and Ing 2005), supporting the observation that the feedback-based dopamine learning signal is time dependent. COVIS provides a strong theoretical basis for RB and II category learning that we can use to interpret the neuroimaging results.

fMRI of RB and II category learning

We implemented a visual category learning experiment in the scanner to examine the neural correlates of RB and II category learning. We also applied mathematical models of category learning to the behavioral data to better characterize the actual strategies being employed by the participants. This modeling then allowed us to reorganize the imaging data in a way that better distinguished the RB and II category learning network activity.

Methods

Participants

Thirty-three healthy, native English-speaking, right-handed adults (10 males) were recruited from the Northwestern University community for participation in this study. All participants gave informed consent according to procedures approved by the Northwestern University Institutional Review Board and were compensated for their time. Participants were randomly assigned to either the RB (N=11) or II (N=22) group. Two II participants were eliminated due to poor quality EPI data (due to subject movement). We recruited twice as many II subjects because initial model analyses revealed a sub-set of II subjects using an RB strategy.

Materials

Stimuli were circular sine wave gratings that varied in spatial frequency and orientation (Maddox, Ashby et al. 2003). Participants were instructed to place each stimulus into one of two categories and to try to learn these categories over time based on the feedback given after each trial. The category structures differed only in the location of the optimal category boundary (Figure 1).

Procedure

On each trial, a fixation cross was presented for 750ms followed by a single stimulus presented for 2 sec. During this time, participants indicated to which category they judged the stimulus belonged. Stimulus offset was followed by a 500 ms visual mask and then 750 ms of feedback ("Right", "Wrong"). A total of 320 categorization trials were performed by each participant in four 80 trial blocks. An equal number of fixation-only trials were pseudo-randomly interspersed between stimulus trials to maximize the separability of the measured hemodynamic response.

MRI acquisition

fMRI data were collected using a GE 3.0 T MRI scanner equipped with a transit/receive head coil while participants performed the categorization task. Whole-brain, gradient-recalled EPI (40 axial 3 mm slices, 0 gap) were collected every 2 sec (TE= 25 ms; flip angle = 78°; 22 cm FOV; 64x64 acquisition matrix; resulting voxel size = 3.44 x 3.44 x 3 mm) for 326 volumes in each of 4 scans. Following the functional runs, high-resolution, 3D MP-RAGE T1-weighted scans (voxel size = 0.859 mm x 0.859 x 1 mm; 160 axial slices) were collected for each participant.

Data analysis

Preprocessing and statistical analysis of the data were performed with a collection of software based on AFNI (Cox 1996). Functional images were co-registered through time to correct for motion, normalized to MNI stereotactic space, and spatially smoothed. Voxels were fit to a general linear model function based on blocking stimuli to track activity that changes as a result of processing specific trials.

Based on a priori hypotheses about the involvement of specific neuroanatomical areas (Nomura, Maddox et al. 2007) we also used a region of interest (ROI) analysis in the hippocampus/parahippocampal gyrus and the caudate nucleus of the basal ganglia. Each individual's ROIs were aligned using the ROI alignment (ROI-AL) method described in Stark and Okada (Stark and Okada 2003).

Mathematical models of RB and II learning

Following previous work (Maddox and Filoteo 2001), two models derived from DBT were fit to each participant's responses to get a more detailed picture of how they were categorizing the stimuli. For each participant, both the RB and II models were fit separately to each of the four 80-trial blocks. The RB model assumed a vertical decision boundary (in stimulus space) reflecting the use of a rule dependent on a single stimulus dimension (e.g. frequency). The II model assumed a decision boundary with slope equal to 1.0 (i.e. a diagonal line reflecting integration of both dimensions). In each case, the model identified the placement of this boundary and the perceptual noise parameter that best accounted for the observed data.

Thus the models both had exactly two free parameters to allow for direct comparison of fit. Best fitting parameter values were identified by a downhill simplex method and the best fitting parameters and corresponding fit value were used to sort the subject runs.

Results and Discussion

Behavioral performance

For both groups of participants, performance was above chance in all runs, and the groups demonstrated similar learning curves. Learning across runs was reflected in a significant linear trend ($F(1, 29) = 24.716$, $p < 0.05$). Mean accuracy averaged across all 4 runs for the RB group was 77.3% (SE = 0.034) and for the II group was 66.6% (SE = 0.055). The RB group accuracy was significantly greater

than the II accuracy across all 4 runs ($F(1,29) = 12.5$, $p < 0.05$).

Modeling behavior

Out of 31 subjects, there were 120 useable blocks of data. The RB model best accounted for 78/120 runs and the II model for 42/120 runs. After sorting the model fits from high to low by group, we selected only the top third from both to be included in the resulting fMRI analyses. In some cases, subjects contributed more than one run to the analysis.

A color-coded schematic of the best fitting model per run is shown in Figure 2. The results of the analysis revealed that the RB subjects, in general were using an RB strategy with the exception of RB subject #11. The II subjects were more heterogeneous. Four out of 20 II subjects were fit consistently with the II model, but five II subjects were fit better with the RB than the II model across all runs. The remaining II subjects were fit better with the RB model on some runs and better with the II model on others, suggesting these people were particularly prone to strategy switching behavior. We used the results of the model analysis to sort the fMRI data in a way that better segregates RB and II strategy use irrespective of the group assignment.

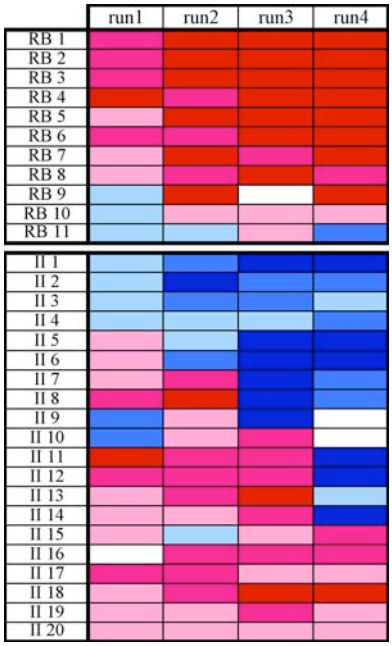

Figure 2: RB and II model fits for RB (top) and II subjects (bottom) across 4 runs. Fit values were separated into thirds with increasing hue indicating better fit. RB fits are in warm and II fits in cool colors. The darkest hues were included in the fMRI analysis. White blocks indicate data loss due to subject movement

fMRI analysis

Region-of-interest (ROI) analyses based on a priori hypotheses in the MTL (Figure 3A) and caudate (Figure 3B) revealed differential activity between groups during successful categorization. The activity patterns in these small volumes are consistent with previous fMRI studies (Nomura, Maddox et al. 2007) and here support the multiple systems view of RB and II category learning.

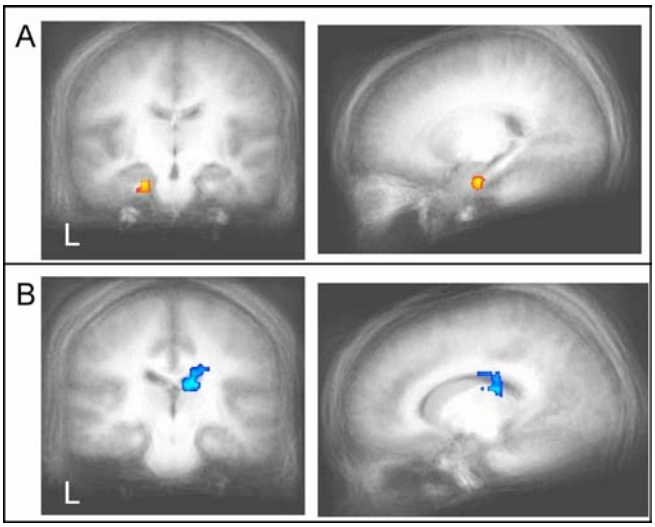

Figure 3: RB vs. II, correct vs. incorrect trials in the (A) MTL and (B) Caudate ROIs. In the left anterior MTL, successful RB categorization activity was greater than that of the II group. In contrast, II successful categorization was greater in the right posterior body of the caudate than RB categorization activity.

Models applied to fMRI data

Mathematical models of RB and II categorization successfully identified subjects within a particular group that were using a non-optimal strategy. In particular, according to the model analysis, the II group contained a large number of subject runs that were better fit with the RB model. This may explain in part the lack of whole-brain group differences in the fMRI analysis above. The models were also useful in that they identified blocks of trials in which a subject was using a strategy particularly well. By sorting the model fits and only examining the functional activity in the top third of these runs, we hoped to isolate activity associated with purely RB or II category learning.

Figure 4A contrasts correct and incorrect trial activity within the top 26 subject runs that were best fit by the RB model. Activity during correct trials was observed in the left anterior MTL and bilateral superior frontal cortex. Similar MTL activity was also seen in the ROI analysis (Figure 3A) using the non-modeled grouping of subjects. The fact that the same area was revealed in this model-based analysis of the same data suggests that the best fit RB runs here may have been driving the effect seen in the ROI. There were also several areas that were more active for these RB-fit subjects when they made an incorrect category judgment, particularly in PFC and medial frontal cortex.

The pattern of activity elicited by correct and incorrect trials suggests that these trials are associated with different cognitive operations. According to COVIS, the RB system generates and selects amongst a variety of rules based on the feedback after each trial. The feedback received during incorrect trials should be an important indication of when the decision boundary needs updating. The data here suggests that the inferior and medial PFC are associated

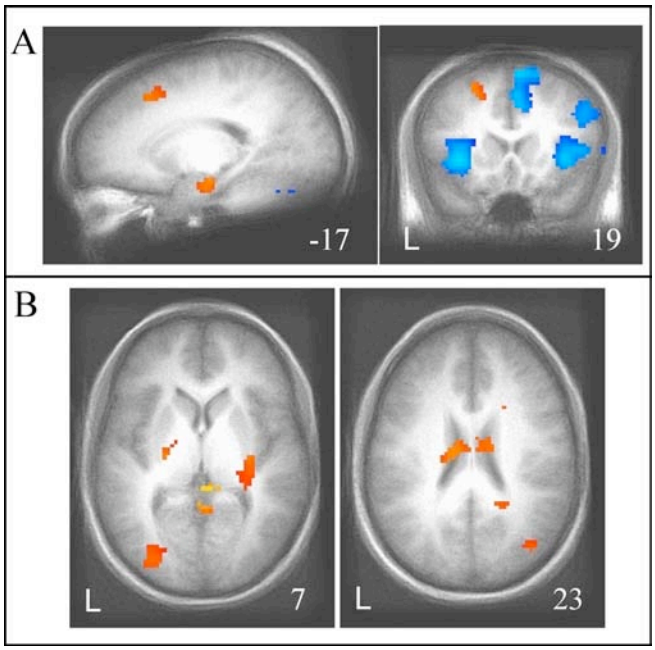

Figure 4: Correct (red) vs. incorrect (blue) trials for (A) best RB model fit runs and (B) best II model fit runs. All images were thresholded at T>4, cluster>350mm^3. When fMRI data was grouped according to model-fit value, activity was detected in areas consistent with the COVIS theory of category learning.

with updating the rule after receiving disconfirming feedback whereas the MTL and frontal cortices are involved with maintaining the appropriate rule. This is one example where fMRI analysis can inform models of categorization. Given two models that make contrasting predictions about feedback processing, the evidence shown here would support a model that assigns the processing of positive and negative feedback to different components of the hypothesized network.

Figure 4B shows the contrast of correct and incorrect trials within the top 14 subject runs that were best fit by the II model. During correct trials, we observed bilateral caudate activity as well as an area in the left visual association cortex. While activity in the caudate was expected based on our ROI analysis, the visual cortical activity was not present in the previous fMRI analysis. In II category learning, the trials in which the subject responded correctly are those in which it is particularly important to link the category label with the appropriate region in perceptual space. As stated in the COVIS theory, one of the hypothesized functions of the caudate is to develop a category representation through the association of activity in posterior caudate with activity in visual association cortex. By only examining functional activity in subjects that were utilizing an II strategy, we've identified a component of this II categorization network that was expected based on a theoretical model.

A specific example of the efficacy of the model in identifying the learning strategy independent of the imposed

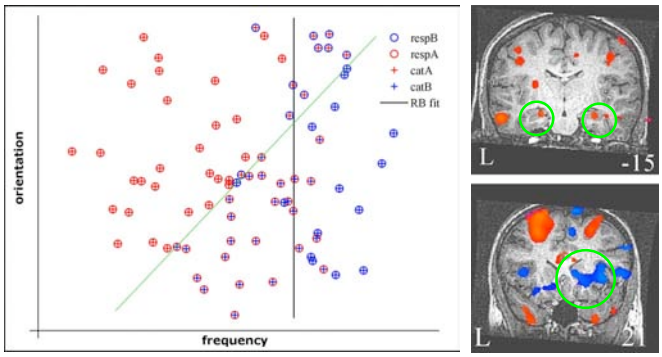

Figure 5: One II subject (#18, run 4) run that was best fit with the RB model (left) and the corresponding contrast of correct vs. incorrect trial activity (right). Plus signs indicate the category membership of the stimuli whereas the circles show the pattern of subject responses. Instances where the colors agree indicate correct responses and all others are incorrect. The green line indicates the optimal decision boundary and the black line the best fitting account of the data according to the RB model. The corresponding functional activity resembles that of the best fit RB run group activity.

category structure is shown in Figure 5. This data comes from a single II subject run that was best fit with the RB model. This subject, despite receiving disconfirming feedback, persisted in using an RB strategy. The pattern of functional activity observed during this run is similar to that of the group of RB-modeled subjects (Figure 4A). Specifically, there is MTL and frontal activity for correct trials and inferior PFC activity for incorrect trials. One could go so far as to suggest that the pattern of activity in the brain could be used to predict the strategy being employed by the individual.

Conclusion

The observed differences in functional activity in the MTL and caudate ROIs suggest that the groups were segregated enough to reveal distinct neural correlates of RB and II category learning as seen previously. That is, although modeling revealed that the II group contained a portion of subjects better fit with an RB model, the sensitive ROI analysis was still able to pull out the activity related to II learning.

The results of the RB and II model-based fMRI analyses revealed a collection of areas that were not detected in the original fMRI analysis. In accordance with the COVIS theory of category learning, the best fitting RB runs were associated with activity in PFC and MTL and best fitting II runs in caudate and posterior visual cortex. The successful utilization of the model-fitting technique shown here to isolate the most effective applications of RB and II strategies suggests that this method in combination with fMRI can be used to further identify the brain networks supporting these processes.

Acknowledgements

The authors would like to acknowledge the funding sources for this research: NIMH R01-MH58748 (PJR), Mechanisms of Aging and Dementia Training Grant T32-AG020418 (EMN).

References

Ashby, F. G., L. A. Alfonso-Reese, et al. (1998). A neuropsychological theory of multiple systems in category learning. Psychol Rev **105**(3): 442-81.

Ashby, F. G. and R. E. Gott (1988). Decision rules in the perception and categorization of multidimensional stimuli. J Exp Psychol Learn Mem Cogn **14**(1): 33-53.

Ashby, F. G. and W. T. Maddox (1990). Integrating information from separable psychological dimensions. J Exp Psychol Hum Percept Perform **16**(3): 598-612.

Ashby, F. G. and W. T. Maddox (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. Journal of Experimental Psychology: Human Perception & Performance **18**: 50-71.

Ashby, F. G. and J. T. Townsend (1986). Varieties of perceptual independence. Psychol Rev **93**(2): 154-79.

Ashby, F. G. and V. V. Valentin (in press). Multiple systems of perceptual category learning: theory and cognitive tests. New York, Elsevier.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res **29**(3): 162-73.

Filoteo, J. V., W. T. Maddox, et al. (2005). Cortical and subcortical brain regions involved in rule-based category learning. Neuroreport **16**(2): 111-5.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. Psychological Review **93**: 411-428.

Knowlton, B. J. and L. R. Squire (1993). The learning of categories: parallel brain systems for item memory and category knowledge. Science **262**(5140): 1747-9.

Lombardi, W. J., P. J. Andreason, et al. (1999). Wisconsin Card Sorting Test performance following head injury: dorsolateral fronto-striatal circuit activity predicts perseveration. J Clin Exp Neuropsychol **21**(1): 2-16.

Maddox, W. T. and F. G. Ashby (1993). Comparing decision bound and exemplar models of categorization. Percept Psychophys **53**(1): 49-70.

Maddox, W. T., F. G. Ashby, et al. (2003). Delayed feedback effects on rule-based and information-integration category learning. J Exp Psychol Learn Mem Cogn **29**(4): 650-62.

Maddox, W. T., F. G. Ashby, et al. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. Mem Cognit **32**(4): 582-91.

Maddox, W. T. and J. V. Filoteo (2001). Striatal contributions to category learning: quantitative modeling of simple linear and complex nonlinear rule learning in patients with Parkinson's disease. J Int Neuropsychol Soc **7**(6): 710-27.

Maddox, W. T. and A. D. Ing (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**(1): 100-7.

Medin, D. L. and M. M. Schaffer (1978). Context theory of classification learning. Psychological Review **5**: 207-238.

Nomura, E. M., W. T. Maddox, et al. (2007). Neural correlates of rule-based and information-integration visual category learning. Cereb Cortex **17**(1): 37-43.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General **115**: 39-57.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. J Exp Psychol Hum Percept Perform **17**(1): 3-27.

Posner, M. I. and S. W. Keele (1968). On the genesis of abstract ideas. Journal of Experimental Psychology **77**: 353-363.

Rao, S. M., J. A. Bobholz, et al. (1997). Functional MRI evidence for subcortical participation in conceptual reasoning skills. Neuroreport **8**(8): 1987-93.

Reed, S. K. (1972). Pattern recognition and categorization. Cognitive Psychology **3**: 382-407.

Smith, J. D., M. J. Murray, et al. (1997). Straight talk about linear separability. Memory and Cognition **23**: 659-680.

Stark, C. E. and Y. Okada (2003). Making memories without trying: medial temporal lobe activity associated with incidental memory formation during recognition. The Journal of Neuroscience **23**: 6748-53.

Waldron, E. M. and F. G. Ashby (2001). The effects of concurrent task interference on category learning: evidence for multiple category learning systems. Psychon Bull Rev **8**(1): 168-76.

Wilson, C. (1995). The contribution of cortical neurons to the firing pattern of striatal spiny neurons. Cambridge, MA, Bradford.