Dissociation of Category-Learning Mechanisms via Brain Potentials

Robert G. Morrison (rmorrison@luc.edu)
Loyola University Chicago, Psychology Department,
6525 North Sheridan Road Chicago, IL 60626 USA

Paul J. Reber (preber@northwestern.edu)
Northwestern University, Psychology Department,
2029 Sheridan Road, Evanston IL 60208 USA

Ken A. Paller (kap@northwestern.edu)
Northwestern University, Psychology Department,
2029 Sheridan Road, Evanston IL 60208 USA

Abstract

Behavioral, neuropsychological, and neuroimaging evidence indicate categories can be learned either via an explicit rule-based mechanism dependent on medial temporal and prefrontal brain regions, or via an implicit information integration mechanism relying on the basal ganglia and occipital cortex. In this study, participants viewed Gabor patches that varied on two dimensions, and learned categories via feedback. Different stimulus distributions can encourage participants to favor explicit rule-based or implicit information integration mechanisms. We monitored brain activity with scalp encephalography while participants (1) passively observed Gabor patches, (2) categorized patches from one distribution, and, one week later, (3) categorized patches from another distribution. Categorization accuracy was matched across the two learning conditions, which nevertheless elicited several distinct event-related potentials. These results demonstrate the efficacy of real-time neural monitoring during category learning and provide additional evidence implicating different neurocognitive mechanisms in explicit rule-based versus implicit information integration category learning.

Keywords: category learning; memory; event-related potentials; ERP; EEG.

Introduction

Behavioral, neuropsychological, and neuroimaging evidence suggests that categories can be learned via explicit and/or implicit mechanisms (Ashby & Maddox, 2005; Kéri, 2003; Nomura & Reber, 2008). Ashby and Maddox (2005) described a feedback category-learning paradigm with different category distributions to selectively encourage one of the two types of learning: Explicit or Rule-Based learning (RB) versus Implicit or Information Integration learning (II). These strategies have been dissociated behaviorally using working memory dual-task procedures (e.g., Zeithamova & Maddox, 2006), feedback delay (e.g., Maddox, Ashby, & Bohil, 2003), and procedural interference (e.g., Ashby, Ell, & Waldron, 2003). In our previous work, we (Nomura et al., 2007; Nomura, Reber, & Maddox, 2007) used functional neuroimaging to demonstrate that RB category learning depends on contributions from prefrontal cortex (PFC) and medial temporal lobe (MTL), whereas II category learning depends on basal ganglia and occipital cortex.

One possible explanation for RB learning specifies a mechanism for hypothesis testing (Ashby et al., 1998). By this account, a participant develops a candidate rule (category A has bars that are thinner than x; category B has bars thicker than x) that is tested based on the feedback on each trial. This RB mechanism would presumably require both updating and maintaining the rule in working memory (dependent on PFC) and updating and maintaining information about the boundary condition in long-term memory (dependent on MTL).

In contrast, II learning appears to occur implicitly, such that the rule for the category structure is difficult or impossible to describe verbally or experience subjectively. II learning may occur via procedural learning under the control of the caudate nucleus in conjunction with visual processing areas in occipital cortex. Dopaminergic reward circuits of the caudate may be responsible for associating specific stimuli with groups of neurons coding for their visual features in occipital cortex (Ashby et al., 1998).

Building on the success of previous neuroimaging

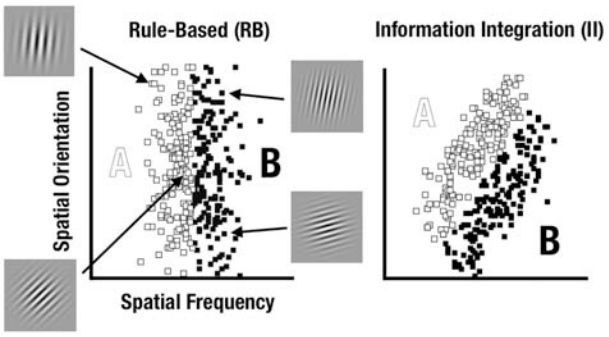

Figure 1. Category distributions used in this study.

efforts to dissociate RB and II category learning, we utilized electrophysiological methods to obtain additional evidence about this neural dissociation with greater temporal precision and to learn more about how the two mechanisms may differ.

Methods

Task Description

We used a visual category-learning paradigm (Maddox, Ashby, & Bohil, 2003) in which subjects learn, via feedback, to categorize Gabor patches that vary in spatial frequency and orientation and that are selected from distributions (Figure 1) designed to encourage either explicit rule-based or implicit information-integration mechanisms.

Participants

Twenty-eight Northwestern University students served as participants in this experiment. Participants received US$15 per hour for two 2- to 3-hour testing sessions. Participants gave informed consent according to the oversight of the Northwestern University Institutional Review Board.

Procedure

Prelearning At the beginning of the first testing session, participants passively viewed 80 RB and 80 II stimuli from the two stimulus distributions to ensure that there were no systematic differences in ERPs based entirely on the two types of distributions. The timing was identical as in the learning trials, but participants made no response and no feedback was given.

Learning In separate sessions, each participant learned an RB category defined by the spatial frequency of Gabor patches and an II category defined by a diagonal threshold based on both spatial frequency and spatial orientation. Sessions were 1 week apart and administered in counterbalanced order. Participants received no instructions about the nature of the categories, but rather discovered the categories with the aid of auditory feedback given 2.5 s after stimulus onset. Figure 2 provides a schematic of the trial timeline. Participants categorized 320 Gabor patches, presented in four blocks, during each learning session. They were debriefed about their categorization strategies after the second testing session.

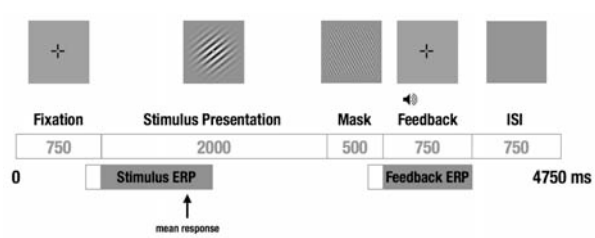

Figure 2. Trial timeline.

EEG Continuous electroencephalographic (EEG) recordings were made during prelearning and learning blocks from 59 evenly distributed scalp sites using tin electrodes embedded in an elastic cap. Four additional channels were used for monitoring horizontal and vertical eye movements. Electrode impedance was ≤ 5 kΩ. EEG signals were amplified with a band pass of 0.05–200 Hz, sampled at a rate of 1000 Hz, and re-referenced offline to average mastoids. Participants were instructed to attempt to refrain from blinking or moving their eye position from fixation during the categorization portion of each trial.

For categorization event-related potentials (ERPs), trials exhibiting eye movements were rejected (< 15% of trials). Averaging epochs lasted 1100 ms, including 100 ms prior to stimulus onset. Participants showed a high of blinking during auditory feedback, so we employed a blink-correction algorithm based on independent component analysis instead of rejecting trials (Source Signal Imaging, 2008). Averaging epochs for feedback processing lasted 850 ms, including 100 ms prior to stimulus onset.

Results

Based on prior behavioral and neuroimaging results, we anticipated that RB and II category learning mechanisms would produce different ERPs particularly when comparing successful (correct) and unsuccessful (incorrect) trials. Specifically, we predicted that ERPs associated with explicit memory (Late Positive Complex/P3) and ERPs associated with selective attention (N1) would show correct-incorrect differences only during RB learning. Given that RB learning is more explicit than II learning, we also anticipated a differential P300 to feedback for incorrect trials in RB versus II learning.

Decision-Bound Theory Modeling

We used mathematical models derived from Decision-Bound Theory (DBT; Ashby, & Maddox, 1993; Nomura et al., 2007) to fit each participant's responses and obtain a detailed picture of how they were likely categorizing the stimuli. The RB model assumed a vertical decision boundary (in stimulus space) reflecting the use of a rule dependent on a single stimulus dimension (spatial frequency). The II model assumed a decision boundary with slope equal to 0.5 (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.

Of the 28 participants in the study, 15 exhibited an II distribution response profile best fit by an II DBT model, while 13 exhibited an II distribution response profile best fit by an RB DBT model (see Figure 3 for

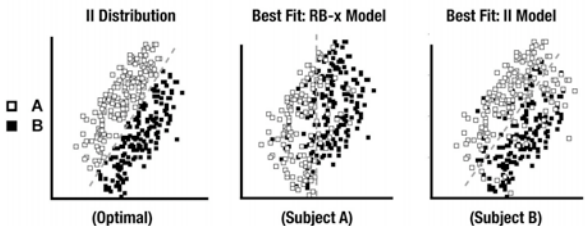

Figure 3. Model fits for two participants in the II condition showing how DBT models can be used to select participants based on likely strategies.

distribution profiles from representative participants). Thus, for analyses reported here we excluded data from participants who were probably using a unidimensional RB strategy to categorize the II category distributions.

Task Performance

Of the 15 participants whose DBT fits were consistent with II strategy use in categorizing the II distributions, 2 did not have an adequate number of incorrect trials to allow for the correct/incorrect ERP analysis (< 30), so their results were also excluded from analysis.

For the remaining 13 participants, accuracy in RB and II conditions (Figure 4) did not reliably differ [$F(1,12) = 1.5$, *ns*]. There was a main effect of block [$F(3,36) = 29$, $p < .001$] and no interaction [$F(3,36) < 1$, *ns*]. Thus, observed differences in correct/incorrect ERP subtractions (described below) cannot easily be attributed to simple differences in accuracy between RB and II learning.

Response time in RB and II conditions did not reliably differ [$F(1,12) = 1.7$, ns]. Correct trials were reliably faster than incorrect trials [F (1,12) = 23, $p <$ .001]. There was no effect of block [F (1,12) = 1, ns]; however, there was a condition by block interaction [F (3,36) = 3.8, $p = .02$] whereby II RTs slightly increased over blocks while RB RTs slightly decreased. Because

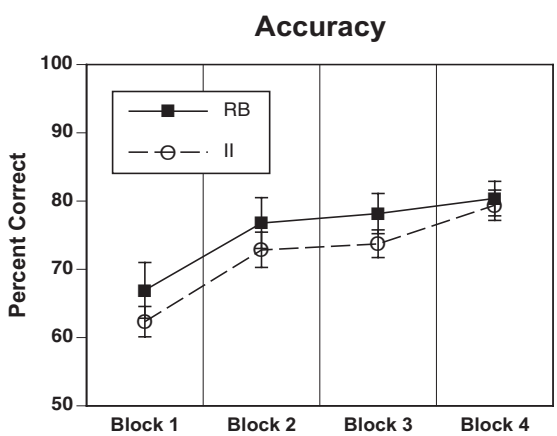

Figure 4. Accuracy results for 13 participants included in ERP analyses.

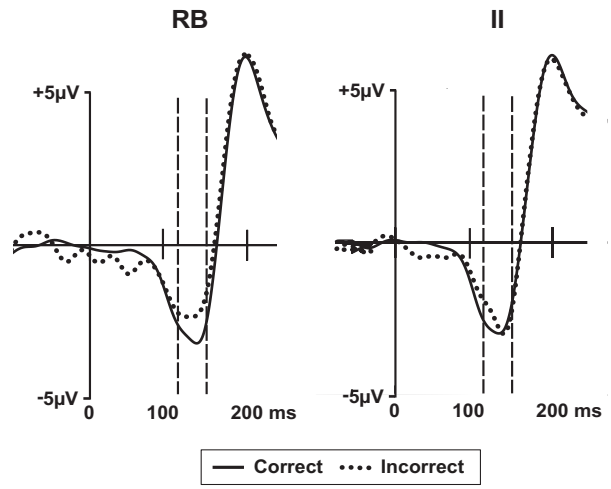

Figure 5. Early negative frontocentral ERPs at 105-155 ms, larger for correct than for incorrect trials during RB learning.

of the difference in correct/incorrect RTs we carefully evaluated the ERPs described in the following sections for onset and offset latencies; however, there was no evidence this differed across accuracy or categorization conditions.

Event Related Potentials

Learning Visual inspection of waveforms and topographies for correct/incorrect subtractions for each categorization condition confirmed three areas of spatiotemporal interest.

First, an early (115-155 ms) negative frontocentral ERP was predictive of correct categorization (Figure 5) in the RB condition [$F(1,12) = 8.1$, $p = .015$], but not the II condition [$F(1,12) = .07$, ns]. Second, a slightly later (170-200 ms) negative occipito-temporal ERP was modulated by categorization condition (Figure 6) with less negativity on correct than incorrect RB trials [$F(1,12) = 8.2$, $p = .014$], but more negativity on correct than incorrect II trials [$F(1,12) = 7.9$, $p = .016$]. Third, a

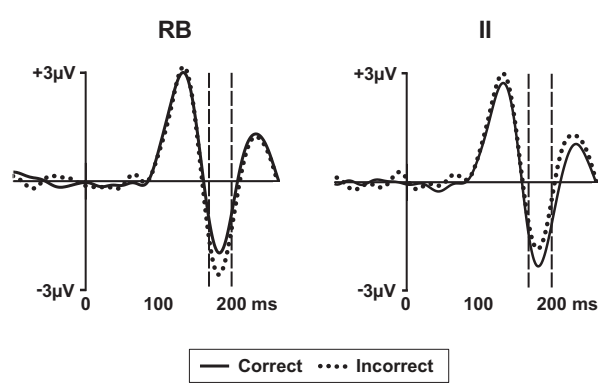

Figure 6. Early negative occipital ERPs differed at 170-200 ms as a function of accuracy and RB/II.

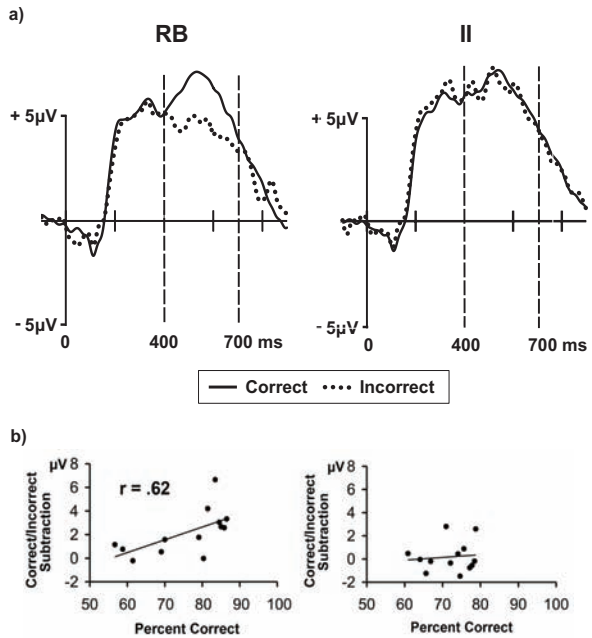

Figure 7. a) Late positive parietal ERPs for RB and II conditions; b) correlations between accuracy and the magnitude of the correct/incorrect subtraction at the peak.

late (400-700 ms) positive parietal ERP was predictive of correct categorization (Figure 7a) in the RB condition [$F(1,12) = 9.0$, $p = .011$], but not in the II condition [$F(1,12) = .077$, ns]. The magnitude of the correct/incorrect subtraction, measured for a 60-ms interval at its maximum, was reliably correlated (Figure 7b) with RB performance [$r(11) = .62$, $p = .02$], but not with II performance [$r(11) = .11$, ns].

Prelearning Although our comparisons of interest during learning were correct/incorrect subtractions performed within a categorization condition (i.e., RB or II), not across conditions, we wanted to ensure that any differences were not due to the nature of the RB or II stimuli distributions. Inspection of prelearning and learning waveforms suggested that early peaks were timed similarly between prelearning and learning, and prelearning peaks were smaller in amplitude than learning peaks. In contrast the prominent late positivity seen in learning was absent in prelearning. We directly compared all three temporal regions of interest during prelearning and found no significant differences between RB and II waveforms at the electrodes that were tested at during learning. This suggests that differences in stimuli per se did not contribute to observed ERP findings during learning.

Feedback In order to assess predictions about the explicit nature of the RB condition relative to the II condition, we examined ERPs recorded during feedback for the presence of a differential P300 response, an ERP sometimes associated with subjective

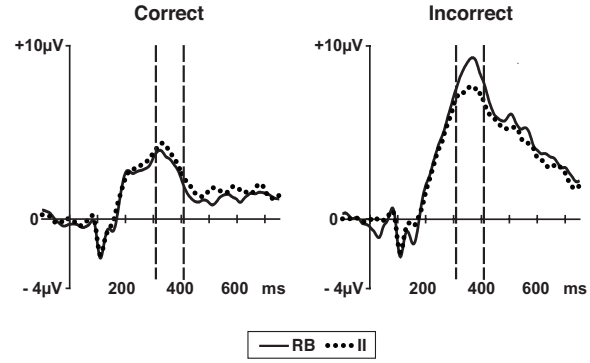

Figure 8. ERPs recorded during feedback show differential RB vs. II P300 responses for incorrect, but not correct trials. Waveform shown is from a right occipital electrode location.

expectations (Polich, 2007). Both correct and incorrect trials showed P300s with broad topographies centered around 350 ms (Figure 8). Whereas ERPs for RB and II conditions did not differ in the 310 to 410 ms range for correct trials [$F(1,12) = .01$, ns], they did differ for incorrect trials [$F(1,12) = 5.7$, $p = .03$], suggesting that participants were more surprised when they learned that they had made an error in the RB condition than in the II condition.

Discussion

Our ERP results provide additional evidence for distinct RB and II category-learning mechanisms. We observed both early and late ERP differences when comparing correct to incorrect trials for RB and II category learning. These ERP differences reflected the distinctive cognitive processing engaged rather than perceptual differences between stimuli or learning conditions.

RB processing is usually thought to depend on hypothesis testing, whereby a candidate rule is evaluated by comparing the representation of the stimulus in the current trial to that of a stimulus representative of the relevant boundary condition (or to some abstract representation of that boundary condition). This evaluation requires selective attention and working memory, likely implemented in prefrontal cortex, as well as the ability to form enduring representations of the rule and boundary condition dependent on hippocampus and medial temporal cortex. In contrast, II learning may be likened to gaining expertise in specialized or holistic processing, as applied for individuating faces or categorizing complex multi-featured objects like Greebles (Rossion, Curran, & Gauthier, 2002).

ERP results were consistent with both of these descriptions. Specifically a differential correct/ incorrect frontocentral N1 ERP (Figure 5) may reflect

early top-down allocation of attentional resources, which is most important in RB learning because one stimulus dimension must be used and the other ignored. In RB learning, the extent to which resources are allocated to the correct feature (i.e., spatial frequency) will tend to result in correct categorization; selectively attending to just one feature in the II condition would often result in errors. Similar frontal N1 potentials have been reported in other visual paradigms (Luck & Vogel, 2000; Hillyard & Anllo-Vento, 1998), and some evidence suggests that the various N1 signals are under frontal control (Deouell & Knight, 2009).

We also observed a differential correct/incorrect response in positive parietal potentials only during RB learning (Figure 7). This effect is thus analogous to the earlier N1 effect. However, this ERP was correlated with performance in the RB condition; subjects with larger correct/incorrect differences performed the task more accurately. Similar positive potentials have been found in many different tasks and variously referred to as the P3, P300, P600 or Late Positive Complex (LPC). These positive potentials have also been associated with working memory (Kok, 2001; Polich, 2007) and episodic memory retrieval (Paller, Voss, & Westerberg, 2009). Thus, LPC potentials found during category learning here may reflect discriminative processing to compare the current Gabor patch with the boundary condition. Differential LPC responses for correct/ incorrect responding in the RB condition are likely to reflect the engagement of the neural system responsible for making the decision based on RB learning.

LPC potentials were also apparent in the II condition with amplitudes for both correct and incorrect trials similar to those for correct RB trials. Importantly, these LPC amplitudes in the II condition were not predictive of accuracy, as they were in the RB condition. This suggests that the neural systems responsible for the LPC may also be engaged during II learning, but are not responsible for the final behavioral decisions. Normura, Reber, and Maddox (2007) argued that RB and II systems compete during categorization and that prefrontal cortex appraises confidence in both systems, making the final decision based on the one with higher confidence for a particular stimuli. Foerde, Knowlton, and Poldrack (2006) likewise demonstrated competition between systems in a categorization task that can also be performed explicitly and implicitly, and were able to experimentally manipulate competition, resulting in changes in the neural systems engaged.

We observed a differential correct/incorrect response in a negative occipitotemporal N1 ERP (Figure 6) in both RB and II conditions. A prior category learning study also revealed differential effects in similarly distributed N1 potentials (Curran, Tanaka, & Weiskopf, 2002). The authors speculated that this ERP could be related to the N170 ERP frequently seen in studies of face processing (e.g., Bentin et al., 1996) and expert categorization (e.g., Rossion et al., 2002; Tanaka & Curran, 2001). This type of processing frequently engages extrastriate visual cortex (e.g., Kanwisher, McDermott, & Chun, 1997; Gauthier et al., 1999), an area found to be more active in the II condition of this task (Normura, Reber, & Maddox, 2007) and previously seen in several other category learning tasks (Reber et al., 1998ab).

One hypothesis is that this N170-like ERP may be sensitive to the type of holistic processing engaged when categorizing complex objects, which could thus overlap with II processing. We found a differential correct/incorrect effect in both II and RB conditions, but importantly, the direction of the effect was inverted; correct II trials showed greater negativity than incorrect trials whereas correct RB trials showed greater negativity than incorrect trials. This pattern of results is consistent with competition between the two systems, such that correct II trials specifically employ holistic processing. We further propose that correct RB trials likely rely on single-feature processing, and incorrect trials may rely more on holistic processing.

Lastly, we observed a differential P300 response comparing RB and II incorrect trials during feedback (Figure 8), with no difference in correct trials. Some researchers have argued that the P300 is an index of cognitive "surprise" (see Polich, 2007). This is consistent with an explicit RB and implicit II mechanism. Specifically, participants in the RB condition are developing firm hypotheses about the rule to use for categorization and the identity of the boundary condition. When those expectations are violated by negative feedback, participants are surprised. In contrast, they are much less certain about what they are doing in the II condition (in spite of equivalent learning as measured by accuracy). This result is also consistent with participants' self reports, which indicate great confidence in rule description after RB learning and little confidence after II learning. These results thus provide further evidence for an explicit/implicit distinction between RB and II learning.

An alternative perspective is that the feedback P300 represents memory updating (see also Polich, 2007). From this perspective, the representation for the boundary condition or the rule must be changed in memory as a result of negative feedback. Again, updating is more important for the system employing explicit memory for categorization, the RB mechanism.

In summary, the present ERP findings illustrate that neurocognitive processes engaged during category learning differ for RB and II learning. These differences occur at multiple time-points in the course of stimulus processing. Real-time neural monitoring via EEG analyses can thereby provide a window into categorization processing yielding information that

goes significantly beyond analyses limited to behavioral responses. Further analyses of these measures may thus constitute an fruitful avenue for gaining new insights into higher cognition generally.

Acknowledgments

We thank Emi Nomura for programming DBT Models, Joel Voss and John Rudoy for technical assistance, Richard Greenblatt and Demetrios Voreades from Source Signal Imaging, and Courtney Clark and Ilya Bendich for assistance in data collection. We are grateful for support from the Northwestern University Mechanisms of Aging and Dementia Training Grant (T32 AG020506, RGM), and the National Science Foundation (0518800 and 0818912, KAP).

References

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

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition, 31,* 1114-1125.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372-400.

Ashby, F.G., Maddox, W.T. (2005) Human category learning. *Annual Review of Psychology, 56,* 149-78.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8,* 551-565.

Curran, T., Tanaka, J.W., & Weiskopf, D.M. (2002). An electrophysiological comparison of visual categorization and recognition memory. *Cognitive, Affective, & Behavioral Neuroscience, 2,* 1- 18.

Deouell, L.Y., & Knight, R.T. (2009). Executive function and higher-order cognition: EEG studies. *Encyclopedia of Neuroscience,* 4, 105-109.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences USA, 103,* 11778-11783.

Gauthier I, Tarr MJ, Anderson AW, Skudlarski P, Gore JC. (1999) Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2,* 568-73.

Hillyard, S. A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences USA, 95*, 781-787.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302-4311.

Kéri, S. (2003). The cognitive neuroscience of category learning. *Brain Research Reviews, 43,* 85-109

Kok A. 2001. On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology 38,* 557–77.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 650-662.

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.

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

Nomura, E. M., Reber, P. J., & Maddox, W. T. (2007). Mathematical models of visual category learning enhance fMRI data analysis. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, Austin TX.

Paller, K.A., Voss, J.L., & Westerberg, C.E. (2009). Investigating the awareness of remembering. *Perspectives on Psychological Science, 4,* 185-199.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology, 118,* 2128-2148.

Reber, P. J., Stark, C. E., & Squire, L. R. (1998a). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory, 5*, 420-428.

Reber, P. J. , Stark, C. E., & Squire, L. R. (1998b). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences USA, 95,* 747-750.

Rossion, B., Gauthier, I., Goffaux, V., Tarr, M. J., & Crommelinck, M. (2002). Expertise training with novel objects leads to left lateralized face-like electrophysiological responses. *Psychological Science, 13,* 250–257.

Source Signal Imaging. (2008). EMSE (Version 5.3) [Computer Software] San Diego, CA. Available from http://www.sourcesignal.com

Tanaka, J. W., and Curran, T. (2001). A neural basis for expert object recognition. *Psychological Science*, *12,* 43–47.

Zeithamova,, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34,* 387-398.