Bridging Levels: Using a Cognitive Model to Connect Brain and Behavior in Category Learning

Todd M. Gureckis (tgurecki@indiana.edu)
Department of Psychology, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Bradley C. Love (love@psy.utexas.edu)
Consortium for Cognition and Computation, University of Texas at Austin
Austin, TX 78712 USA

Abstract

Mental localization efforts tend to stress the where more than the what. We argue that the proper targets for localization are well-specified cognitive models. We make this case by relating an existing cognitive model of category learning to a learning circuit involving the hippocampus, perirhinal, and prefrontal cortex. Results from groups varying in function along this circuit (e.g., infants, amnesics, older adults) are successfully simulated by reducing the model's ability to form new clusters in response to surprising events, such as an error in supervised learning or an unfamiliar stimulus in unsupervised learning. Reported task dissociations (e.g., categorization vs. recognition) are explained in terms of cluster recruitment demands.

A major goal of cognitive psychology has been to develop an understanding of behavior in terms of computational principles. However, we are often left with the question of what these models tell us about the brain. The answer is certainly not clear. The growing area of cognitive neuroscience offers an endless source of new embers for this debate, as more and more cognitive function is localized and described in terms of specific brain processes. However, by focusing on the localization of mental function (i.e., where is processes X in the brain?), we run the risk of amassing a list of brain areas associated with certain tasks in the absence of useful linking theories reflecting how those regions interact to control behavior in our daily lives.

In this paper, we argue that well-specified, process models of cognitive functions are the appropriate targets for localization. Successful process models offer a number advantages over folk psychological, ad hoc, or traditional psychological theories. For example, model developed in cognitive psychology make predictions, have mechanisms and dynamics which can be related to brain measures, and offer a simple and clear starting point for developing theories of brain function. To support our conjecture, we focus on relating a process model of human category learning to a learning circuit involving the hippocampus, perirhinal cortext, and prefrontal cortex (PFC).

The model we consider, Supervised and Unsupervised STratified Adaptive Incremental Network (SUSTAIN), is applied to human data from a number of populations (infants, amnesics, and older adults) who differ in their category learning ability. Armed with its computational principles and the proposed mapping, SUSTAIN is able to predict how degraded function along this circuit affects category learning performance for these (and other) groups. In particular, SUSTAIN relates the degree of preserved function to how readily members of a group can individuate events, as opposed to collapsing experiences together into a common gestalt (see Figure 1).

After introducing the model, we explain the close correspondence between aspects of the model and the currently understood function of a learning circuit involving PFC, the hippocampus, and perirhinal cortex. We then review a number of simulations which support our theory. In doing so, we provide a novel framework for understanding the role this circuit plays in category learning ability. In addition, our analysis suggests a recasting of several dichotomies popular in the field, such as the distinction between categorization and recognition, recollective and familiarity-driven responding, and episodic and semantic memory.

SUSTAIN and the Proposed Mapping

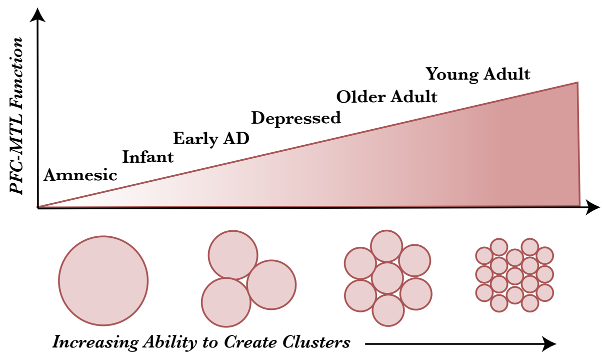
We begin by introducing the computational theory (SUSTAIN) and the bridge theory linking SUSTAIN to functional components in the brain. Due to limited space, readers interested in the mathematical details of the model are directed elsewhere (Love, Medin, & Gureckis, 2004).

SUSTAIN

The basis for representing category knowledge has been proposed to be rule-based, exemplar-based, or prototype-based. SUSTAIN proposes that clusters, which display characteristics of all three of the aforementioned approaches, underlie our category representations. A cluster is a bundle of features that captures conjunctive relationships across dimensions. For example, a cluster can capture the fact that having wings, flying, and having feathers tend to co-occur.

In SUSTAIN, categories are represented by one or more clusters. For example, the category of birds might be represented by multiple clusters which capture natural patterns of regularity within the category (e.g., song birds, birds of prey, penguins, and ostriches each might be represented by separate clusters which all belong to the super-ordinate). A cluster can also belong to multiple categories at once because they are linked to categories by association weights that are adjusted during learning.

Figure 1: Various groups are ordered by PFC-MTL function. Groups with higher PFC-MTL function are predicted to have an increasing ability to individuate items in memory (i.e. create new, distinct clusters), while low functioning groups such as amnesics are assumed to collapse items in a single gestalt. SUSTAIN captures differences along this continuum by varying a single parameter related to how readily additional clusters are recruited during category learning.



SUSTAIN's clusters mediate the relationship between inputs (e.g., stimulus presentation) and output (e.g., category assignment). SUSTAIN begins with one cluster centered on the first training item. Additional clusters are recruited in response to surprising events. In unsupervised learning, a surprising event is exposure to a sufficiently unfamiliar or novel stimulus. In supervised learning, a surprising event is a classification error. SUSTAIN's recruitment scheme implies that surprise drives differentiation of critical stimulus patterns. When a surprising event does not occur, the current stimulus is assigned to the dominant cluster (i.e., the cluster most activated or similar to the current item) and this dominant cluster moves towards the current stimulus so that the cluster converges to the centroid or prototype of its members. Thus, in the absence of surprise, events are collapsed together in memory. The ability to create new clusters forms a continuum. At one extreme, SUSTAIN is a prototype model where each category is represented by a single cluster, while at the other extreme the model becomes an exemplar model where each item is captured in its own cluster.

Mapping Hypothesis

Central to SUSTAIN is the ability to form new clusters in response to surprising events. This type of learning is rapid and involves forming episodic codes or traces which support subsequent learning. A mature and intact learning circuit involving the hippocampus, PFC, and perirhinal cortex is assumed to underly this ability. At its essence, our theory predicts how degraded function in this learning circuit affects learning performance for various populations. Groups with higher PFC-MTL function are predicted to have an increased ability to individuate items in memory (i.e. create new, distinct clusters), while low functioning groups such as amnesics are assumed to collapse items in a single gestalt.

Figure 2: The Logic of Cluster Recruitment.

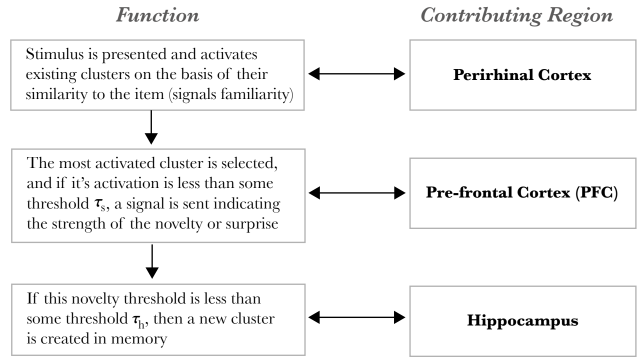


Figure 2 describes the logic of cluster recruitment and highlights the nature of the mapping between this process and components of the PFC-MTL circuit. Newly formed clusters are assumed to be created by the hippocampus. These codes migrate to structures in the Medial Temporal Lobe (MTL), such as the perirhinal cortex, through reciprocal connections with entorhinal cortex. These existing representations in the perirhinal cortex support a signal of familiarity or fit. The PFC is assumed to monitor surprise and direct encoding and retrieval processes.[1]

In support of our proposal, PFC and perirhinal cortex are interconnected and participate in a circuit that could direct the hippocampus's encoding of surprising events (see Ranganath and Rainer, 2003, for a review). The PFC monitors surprise by comparing the current stimulus to representations in perirhinal cortex (which provides a measure of familiarity or fit) and based on this comparison directs hippocampal encoding. Indeed, lesioning the connection between PFC and perirhinal cortex eliminates the memory advantage for surprising items (Parker, Wilding, & Akerman, 1998). In ERP studies, the PFC is associated with the P3 novelty signal which orients attention towards novel stimuli and which has been found to correlate with item memory (Ranganath & Rainer, 2003).

The key component of the theory brought out in the subsequent simulations concerns the idea that the creation of new codes (or clusters) depends on a intact and mature hippocampus. Broadly constructed, clusters capture key patterns and relations between stimulus elements into a discrete bundle. In this way, cluster creation in SUSTAIN is akin to the creation of a "conjunctive code" (a representation thought to be dependent on the hippocampal formation) (Sutherland, McDonald, Hill, & Rudy, 1989). Findings from amnesic patients with hip-

[1]Note that disruption anywhere along the learning circuit can result in the failure to encode a surprising event. In modeling terms, the degree of PFC and hippocampal function are captured by separate parameters. Here, we focus on populations and tasks in which hippocampal function should be the limiting factor.

pocampal lesions support this position. For example, amnesic patients equated with controls for item recognition remain impaired in list discrimination (Downes, Mayes, MacDonald, C., & Hunkin, 2002), which requires encoding the conjunctions of item and list, and patients with hippocampal damage make more conjunctive errors than healthy controls (Kroll, Knight, Metcalfe, Wolf, & Tulving, 1996).

In SUSTAIN, a single parameter relates to the degree of spared hippocampal function. At one extreme are amnesic patients lacking a hippocampus at the other are young normals. The parameter controls SUSTAIN's ability to form clusters that are similar to existing clusters. When the parameter is set low (poor hippocampal function), SUSTAIN can only successfully form a new cluster when the current stimulus is drastically different from any existing cluster causing most events or episodes to be undifferentiated.

Amnesics, Category Learning, and Recognition Memory

As shown in Figure 1, our theory characterizes amnesic patients by their inability to individuate events (i.e., to recruit clusters in response to surprising events). Knowlton and Squire's (1993) studies illustrate this point in a category learning task. K&S found that amnesic patients can categorize, but not recognize, dot pattern stimuli at accuracy levels comparable to matched controls (but see Palmeri and Flanery, 1999).

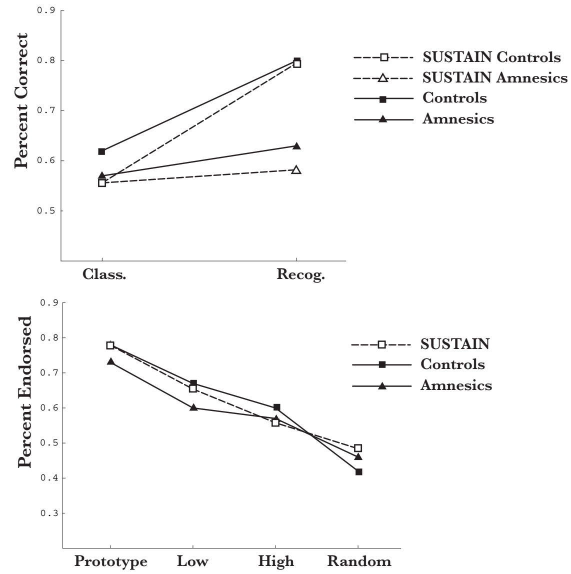
In K&S's categorization task, participants viewed twenty low and twenty high distortions of an underlying prototype during the study phase. Participants were then informed that these items all belonged to a common category. At test, participants indicated whether the presented stimulus was a member of the category. Half the test trials consisted of random pattern unrelated to the prototype underlying the study items. The other half of the test trials included the presentation of the prototype (which was not actually shown during the study phase), novel low distortions of the prototype, and novel high distortions of the prototype.

In the recognition task, participants viewed five distinct patterns eight times each. At test, participants were shown these five items and five random foils and indicated whether the stimulus was shown in the study phase. The main results, illustrating a dissociation between categorization and recognition performance for amnesic patients, along with SUSTAIN's fit of these data, are shown in Figure 3 (top). In the categorization test phase, both groups (and SUSTAIN) displayed a generalization gradient that fell as similarity to the prototype decreased (Figure 3, bottom).

Simulations of Knowlton and Squire (1993)

Low hippocampal function in SUSTAIN is modeled by reducing the model's ability to recruit a clusters in response to a surprising event in the presence of similar, existing clusters. Thus, groups low in hippocampal function are modeled as having a lower setting of the threshold parameter which controls this ability. Besides differences along this single parameter, simulations of amnesics and controls were identical.

Figure 3: Top: The main results from Knowlton and Squire (1993) are shown along with SUSTAIN's fit. Bottom: The generalization gradient for participants in the categorization test phase along with SUSTAIN 's fit



In K&S's categorization task, simulations for both groups recruited only a single cluster. The nature of the study items are sufficiently similar to one another that they were all collapsed into a single cluster. Because both groups have the same internal representation of the study items (i.e., one cluster), SUSTAIN necessarily predicts equivalent performance for the two groups.

In contrast, the simulations for the controls in the recognition task result in five clusters being recruited. The five distinct patterns shown in the recognition study phase are sufficiently dissimilar that they are individuated (i.e., each item is surprising when initially presented). In the amnesic simulations of the recognition task, each item is also surprising when initially presented, but because of low hippocampal function, clusters are not always recruited in response to these surprising events. Instead of recruiting five clusters as in the normal simulations, SUSTAIN recruited 2-4 clusters in the amnesic simulations collapsing either some or all of the items together in a single cluster.

Multiple Systems?

K&S interpret their results in terms of multiple learning systems engaged in categorization and recognition. Following Nosofsky and Zaki (1998), we suggest that recognition and categorization are essentially equivalent. In fact, in our modeling, recognition and categorization are modeled in an identical fashion. Our results suggest that the underlying variable for explaining performance differences between the two populations is not recognition or categorization, but reliance on conjunctive codes (i.e., number of clusters required). In the case of K&S's

design, only one cluster is required for successful categorization and predicted performance is identical for both groups. However, for a more difficult category learning task that requires multiple clusters (e.g., one that involves multiple categories or category subtypes), SUSTAIN predicts amnesic patients should show a deficit relative to controls (in line with Zaki's 2004 meta-analysis).

The simulations reported here bear a resemblance to Nosofsky and Zaki's (1998) modeling of K&S (1993) with an exemplar model. In their account, amnesic patients were modeled as having broader generalization gradients around each exemplar relative than controls. This led to a "blurring" or averaging of the internal representations for simulated amnesics, which is quite similar to the clustering process we advocate here. However, our account differs in that we place the locus of deficit at encoding rather than at recall. Finally, while Nosofsky and Zaki advocated a single-system view of recognition and categorization, we see it as likely that multiple and overlapping systems contribute to both recognition and categorization.

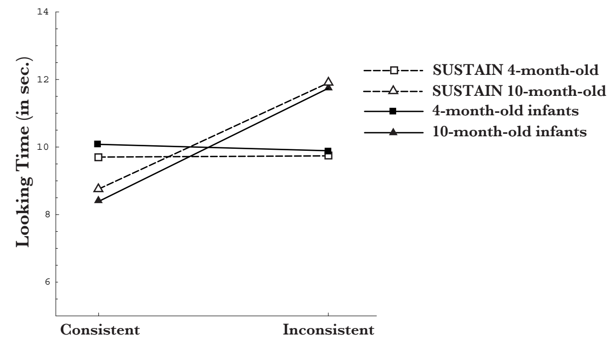
Category Learning in Early Infancy

Considerable developmental changes occur in the hippocampus during the first year of life. For example, hippocampal volume reaches adult level during second half of first year of life (Kretschmann, Kammradt, Krauthausen, Sauer, & Wingert, 1986), cell differentiation surges in the hippocampus between 7-10 months (Seress & Mrzljak, 1992) and continues into the second year while connectivity between the hippocampus and other areas continues to increases into the second year (Benes, 1994). These changes are consistent with a general "back to front" trajectory in neurological development, starting with basic visual areas and progressing toward frontal areas (Johnson, 2003).

Given these developmental changes, one reasonable hypothesis is that infants' PFC-MTL learning circuit will not be fully functional. In terms of the continuum shown in Figure 1, young infants are assumed to have ability closer to amnesic patients than to young adults. Thus, infants should show difficulty in conjunctive tasks that require multiple clusters.

Younger and Cohen (1986) conducted a series of studies assessing developmental changes in category learning ability. In their studies, both four and ten-month-old infants were habituated to a set of sequentially presented visual stimuli depicting imaginary animals which varied along three attributes. Two of these attributes correlated perfectly across the habituation items. The abstract structure of these items was [1 1 1], [1 1 2], [2 2 1], and [2 2 2] where the first two dimensions were correlated (i.e., a particular type of animal head always appeared with a particular type of tail). After the habituation phase, infants were shown test items that either followed the correlated pattern of the habituation items or violated it (i.e., the consistent test item was [2 2 2] and the inconsistent item was [2 1 1]). If infants learned that nature of the relationship between the correlated dimensions, they should find the uncorrelated item to be

Figure 4: Looking times for four-month-old and ten-month-old infants in Younger and Cohen's (1986) Experiment 2 are shown along with SUSTAIN's fit.



novel and therefore look at it longer. In contrast, if infants failed to encode the attribute relation, but instead only encoded attribute-value (i.e., feature) frequencies, both test items should be equally interesting and should yield equal looking times.

The basic findings, are shown in Figure 4.[2] Four-month-old infants looking times for the consistent and inconsistent items were equal, whereas ten-month-old infants devoted more time to the inconsistent item than to the consistent item.

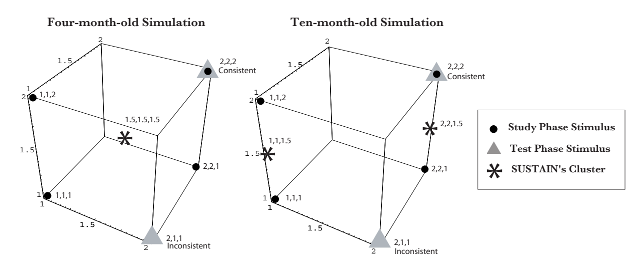
Simulations of Younger and Cohen (1986)

Due to the continued maturational changes in PFC-MTL function four-month-old infants were modeled with a lower setting of the hippocampal function parameter (like amnesics). SUSTAIN displayed the same pattern of results as the human infants (see Figure 4). Figure 5 shows the spatial configuration of SUSTAIN's clusters relative to study and test items. In the four-month-old simulations (shown in the left plot of Figure 5), SUSTAIN recruited a single cluster. The single cluster represents the average of the four study items (i.e., the feature frequencies). This single cluster is located in the center of the space and is equidistant from both the consistent and inconsistent test items. Given this configuration, SUSTAIN predicts that both test items are equally familiar.

In the 10-month-old simulations (shown in the right plot of Figure 5), SUSTAIN recruited two clusters. Each of these two clusters represented the average of two of the four study stimuli. One cluster represented the average of stimuli [1 1 2] and [1 1 1] (located at 1 1 1.5), whereas the other represented the average of stimuli [2 2 2] and [2 2 1] (located at 2 2 1.5). In this case, the inconsistent test stimulus is farther from the nearest cluster than the consistent test stimulus is. This effect is magnified by SUSTAIN's shift of attention to the two correlation-relevant attributes. These two clusters effec-

[2]We focus here on the results of Experiment 3, although SUSTAIN successfully fits all the studies contained in Younger and Cohen where infants showed learning (Gureckis & Love, 2004).

Figure 5: A geometric rendering of the stimulus structure from Exp. 2 along with SUSTAIN's clustering solutions is shown.



tively encode the conjunctive relationship between the first two attributes of the study items.

These simulations closely parallel the simulations of amnesics in K&S (1993). Both young infants and amnesic patients are sensitive to feature frequency, but not to feature relations.

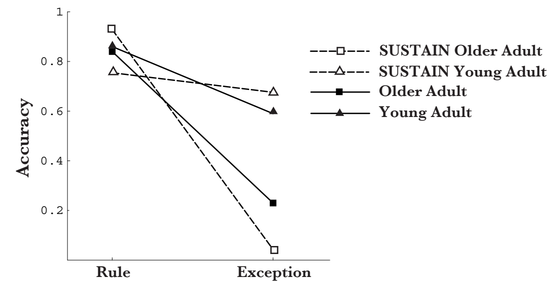
Effect of Aging on Category Learning

Aging does not affect the brain uniformly. Imaging and neuroanatomical studies reveal greater shrinkage in the hippocampus and PFC compared to other areas, such as parietal and occipital cortex (Flood & Coleman, 1988; Raz, 2000). Released in response to environmental stress over the course of our lives, cortisol has been shown to reduced hippocampal volume and is linked to deficits in hippocampal mediated memory tasks (Lupien et al., 1998). Given these assaults on the PFC-MTL learning circuit, we predict that (like infants and amnesics) older adults should be impaired at conjunctive learning tasks which (in terms of SUSTAIN) require the multiple clusters.

Simulating the effects of aging on category learning

Love, Gureckis & Worchel (under review) report a study aimed at directly comparing younger and older adult performance in a single task. Younger and older adults were trained by supervised classification learning on two contrastive categories. Each category was defined by an imperfect rule, such that all members of the category shared a value along a single attribute except for a single exception. For instance, if the critical dimension was size, then all members of category A might be "large" except for the one exception which would be "small". Likewise, all members of category B would be "small" except for a single "large" exception. Participants completed 80 study trials and then completed a test phase consisting of the eight studied items and eight novel items that contained the same features as the studied items re-arranged. In the test phase, participants indicated the category membership of the stimulus as in the study phase.

Figure 6: Study phase accuracy for the study phase of the aging study are shown, along with SUSTAIN's fit.



The main results from the study phase are shown in Figure 6. Older adults performed equivalently to younger adults on rule-following items, but showed a large deficit on the exception items. Within the older adult group, the difficulty with exception items increased with age, whereas performance actually increased for rule-following items with age. SUSTAIN's predictions mirror this result.

When modeling older adults, SUSTAIN's hippocampal parameter was set low in order to limit the model's ability to form new clusters that are similar to existing clusters (as in previous simulations of amnesic patients and four-month-old infants). SUSTAIN predicts that exception items can only be mastered by forming separate clusters to encode these items. Because exception items will be fairly similar to existing clusters that capture rule-following items from the opposing category, SUSTAIN predicts that these items should be especially difficult for older adults to master (as shown in Figure 6).

For the younger adults, SUSTAIN created separate clusters for each exception allowing it to eventually master these items. In contrast, in the older adult simulations, SUSTAIN assigned the exception items to clusters that capture the rule-following items from the *opposing* category. For these simulations, SUSTAIN failed to individuate the exception items and treated these items as if they provided support for the discriminative rule. As a result, SUSTAIN's clusterings for the older adult simulations predict that older adults should have more abstract rule representations than younger adults (their solution of collapsing all rule-following items into a single cluster for each category does not perserve item-specific information and instead stresses the rule dimension).

This prediction held. Older adults applied the imperfect rule in the transfer phase as often to rule-following items that appeared in the study phase as they did novel items. In contrast, younger adults rule application was more influenced by similarity to exemplars seen in the study phase, a pattern of results which prove challenging for exemplar theories (Nosofsky & Zaki, 1998).

Discussion

In this paper, we argue that localizing cognitive models (that simulate interesting behaviors) are our best bet for directing and understanding empirical research. At some level, every researcher relies on a model of how cognition works, even if that model is not explicitly acknowledged. We argue the best model to use is one that is well-specified, relatively simple, and verified empirically, that is, an existing cognitive model.

We demonstrated this by applying an existing model of category learning to category learning results from hippocampal amnesics, young infants, young adults, and older adults. The mapping between the model and the PFC, hippocampus, and perirhinal cortex was simple and incomplete, yet was sufficiently powerful to place findings from these numerous subfields into a common theoretical framework. Our account provides new perspective on the influence of hippocampal impairment on category learning which could have implications for both treatment and diagnosis. Although not addressed here, our theory also applies to other groups such as clinically depressed, those suffering from the early stages of Alzheimer's, or even to groups at the opposite extreme who might hyper-individuate such as autistics.

In addition, our account provides an unique perspective on many popular debates in the field. For example, our simulations suggest that it is the number of distinct codes needed to learn a task which is the critical dimension influencing learning rather than arbitrary task labels such as "categorization" or "recognition". Our theory naturally predicts that differences between groups in their ability to individuate events in memory interact with task demands to explain performance. SUSTAIN's cluster recruitment method also blurs the distinction between semantic and episodic memory. New clusters begin as distinct, episodic traces encoding exceptions and surprisingly novel stimuli, but later may evolve to be more abstract through continued learning making such distinctions one of degree rather than of kind.

Acknowledgments

This work was supported by AFOSR grant FA9550-04-1-0226 and NSF CAREER grant #0349101 to B.C. Love and NIH-NIMH training grant #:T32 MH019879-12 to T.M. Gureckis. Correspondence concerning this research should be addressed to either author.

References

Benes, F. (1994). Cortico-limbic system. In G. Dawson & K. Fisher (Eds.), *Human behavior and the developing brain* (p. 176-206). New York: Guilford Press.

Downes, J. J., Mayes, A. R., MacDonald, C., & Hunkin, N. M. (2002). Temporal order memory in patients with korsakoff's syndrome and medial temporal lobe amenesia. *Neuropsychologia*, *40*, 853-861.

Flood, D. G., & Coleman, P. D. (1988). Neuron numbers and sizes in aging brains: Comparisons of humans, monkeys, and rodent data. *Neurobiology of Aging*, *9*, 453-463.

Gureckis, T., & Love, B. C. (2004). Common mechanisms in infant and adult category learning. *Infancy*, *5*, 173-198.

Johnson, M. H. (2003). Development of human brain functions. *Biological Psychiatry*, *54*, 1312-1316.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*, 1747-1749.

Kretschmann, J. J., Kammradt, G., Krauthausen, I., Sauer, B., & Wingert, F. (1986). Growth of the hippocampal formation in man. *Bibliotheca Anatomica*, *28*, 27-52.

Kroll, N. E. A., Knight, R. T., Metcalfe, J., Wolf, E. S., & Tulving, E. (1996). Cohesion failure as a source of memory illusions. *Journal of Memory and Language*, *35*, 176-196.

Love, B. C., Gureckis, T. M., & Worchel, J. (submitted). Category learning and aging: No exception is the rule.

Love, B. C., Medin, D. L., & Gureckis, T. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, *111*, 309-332.

Lupien, S. L., de Leon, M., de Santi, S., Convit, A., Tarshish, C., Nair, N. P. V., Thakur, M., McEwen, B. S., Hauger, R. L., & Meaney, M. J. (1998). Cortisol levels during human aging predict hippocampal atrophy and memory deficits. *Nature Neuroscience*, *1*(1), 69-73.

Nosofsky, R. M., & Zaki, S. F. (1998). Dissociations between categorization and recognition in amnesic and normal individuals. *Psychological Science*, *9*, 247-255.

Palmeri, T. J., & Flanery, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, *10*, 526-530.

Parker, A., Wilding, E. L., & Akerman, C. (1998). The von Restorff effect in visual object recognition memory in humans and monkeys: The role of frontal/perirhinal interaction. *Journal of Cognitive Neuroscience*, *10*, 691-703.

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, *4*(3), 193-202.

Raz, N. (2000). Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In F. I. M. Craik & T. A. Salthouse (Eds.), *Handbook of aging and cognition* (Vol. 2, p. 1-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Seress, L., & Mrzljak, L. (1992). Postnatal development of mossy cells in the human dentate gyrus: A light microscopic golgi study. *Hippocampus*, *2*, 127-142.

Sutherland, R. J., McDonald, R. J., Hill, C. R., & Rudy, J. W. (1989). Damage to the hippocampal formation in rats selectively impairs the ability to learn cue relationships. *Behavioral and Neural Biology*, *52*, 331-356.

Younger, B., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations amoung attributes. *Child Development*, *57*, 803-815.

Zaki, S. (2004). Is categorization performance really intact in amnesia? A meta-analysis. *Psychonomic Bulletin & Review*, *11*, 1048-1054.