Hippocampal-System Function in Stimulus Representation and Generalization: A Computational Theory

Mark A. Gluck **Catherine E. Myers**

Center for Molecular and Behavioral Neuroscience
Rutgers University
197 University Ave., Newark, NJ 07102
gluck@pavlov.rutgers.edu *myers@pavlov.rutgers.edu*

Abstract

We propose a computational theory of hippocampal-system function in mediating stimulus representation in associative learning. A connectionist model based on this theory is described here, in which the hippocampal system develops new and adaptive stimulus representations which are predictive, distributed, and compressed; other cortical and cerebellar modules are presumed to use these hippocampal representations to recode their own stimulus representations. This computational theory can been seen as an extension and/or refinement of several prior characterizations of hippocampal function, including theories of chunking, stimulus selection, cue-configuration, and contextual coding. The theory does not address temporal aspects of hippocampal function. Simulations of the intact and lesioned model provide an account of data on diverse effects of hippocampal-region lesions, including simple discrimination learning, sensory preconditioning, reversal training, latent inhibition, contextual shifts, and configural learning. Potential implications of this theory for understanding human declarative memory, temporal processing, and neural mechanisms are briefly discussed.

Introduction

The hippocampus and adjacent cortical regions in the medial temporal lobe have long been implicated in learning and memory via lesion data in both humans (Scoville & Millner, 1957; Squire, 1987) and animals (Mishkin, 1982; Squire and Zola-Morgan, 1983). While there is general agreement that this region plays an essential role in many aspects of learning and memory, there is little consensus as to the precise specification of this role. One approach has been to seek to define the class of learning and memory tasks which require an intact hippocampal region. Squire (1987) emphasizes the critical role of this brain region for the formation of explicit declarative memories in humans. Studies of lower (non-primate) mammals have focussed on place-learning and spatial navigation as tasks which require an intact hippocampal region (Morris, et al., 1982; O'Keefe and Nadel, 1978; McNaughton & Nadel, 1990).

Another approach to functional theories of hippocampal-region processing has been to characterize an underlying information-processing role for the hippocampal region and then seek to derive a wider range of task-specific deficits. Two broad classes of hippocampal-based deficits have been characterized: those dealing with temporal processing and those concerned with modifying stimulus representations. We are concerned here solely with the latter, representational, properties of hippocampal-region function. Thus, in the analysis to follow, we will consider only trial-level properties of associative learning.

Several different representational theories of hippocampal-region function have been proposed. Wickelgren (1979) suggested that the hippocampus participates in a process whereby the component features within a stimulus pattern are recognized as co-occurring elements and thus come to be treated as a unitary whole or "chunk." Others have viewed the hippocampal-region as an attentional control mechanism which alters stimulus selection through a process of inhibiting orienting or attentional responses to irrelevant cues (Douglas & Pribram, 1966; Schmajuk & Moore, 1985). More recently, Wickelgren's "chunking" idea has been extended and elaborated by Sutherland and Rudy (1989) who proposed that the hippocampus provides the neural basis for the acquisition and storage of configural events, while other brain systems store only direct cue-outcome associations. A related characterization of hippocampal-function suggests that it is best understood as provid-

ing a "contextual tag" for associative learning (Nadel & Willner, 1989; Winocur, et al., 1987; Hirsch, 1974). Eichenbaum and colleagues have also emphasized the representational role of hippocampal function, particularly in the flexible use of conjunctive associations in novel situations (Eichenbaum & Buckingham, 1991).

These characterizations of hippocampal-lesion deficits can all qualitatively describe subsets of the empirical data. What is lacking, however, is a clear mechanistic interpretation of hippocampal-region function which can be formally and rigorously tested against a broad range of empirical data. Computational models of abstract connectionist networks offer one possible approach for exploring candidate functional roles of the hippocampal system. Models developed at this abstract level do not directly yield a physiological understanding of circuit-level processing in the hippocampal region; nevertheless, these models may suggest how effects of hippocampal-region lesions might emerge from an underlying processing function which is localized in the hippocampal region. By developing a connectionist theory of hippocampal processing in conditioning, we seek, in the work to be described here, to address the question: Is there a simple underlying *computational* function which can derive the representational processes subserved by the hippocampal system in associative learning?

Cortical-Hippocampal Model

We begin with the single key idea that the representational function of the hippocampal system can be approximated by a simple network architecture, called a predictive autoencoder. This type of network develops novel and flexible representations with three key properties: they are **distributed**, **predictive** (of future sensory inputs), and **compressed** (i.e., reduced in size by compressing statistical redundancies). In our model, the hippocampal module is conceptualized as a predictive autoencoding network, which learns sensory-sensory mappings through a narrow hidden layer, as shown in Figure 1 (c.f., Hinton, 1989). As a result, the narrow hidden layer is forced to discover a representation which compresses regularities and irrelevant stimuli while allocating more resources to predictive stimuli. Other learning "modules" such as cerebellar and cerebral cortices (one such module is shown in Figure 1) are restricted to learn simple associations via procedures such as the Rescorla-Wagner rule (Rescorla & Wagner, 1972). If, however, a linear recombination of the hippocampal module's

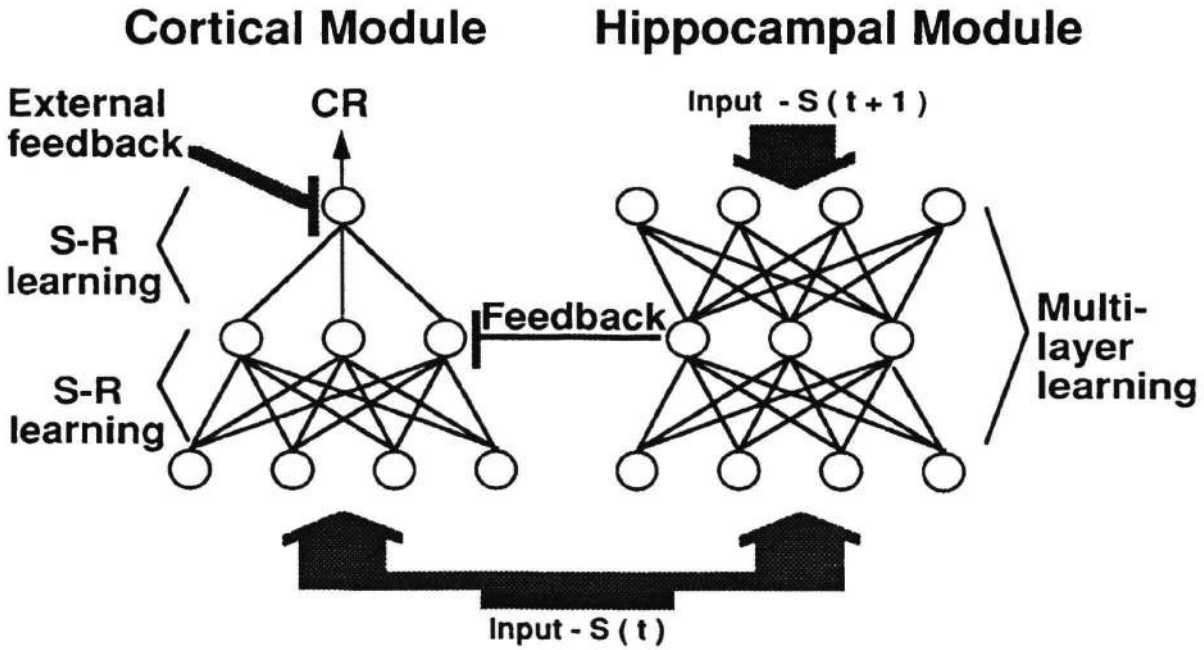

Figure 1. *The intact cortico-hippocampal model. Learning in the hippocampal module (on right) mediates the development of novel stimulus representations in cerebral and cerebellar modules (one shown on left). The hippocampal module has the capacity for multi-layer learning, which results in a novel recoding (or re-representation) of its stimulus inputs. The cortical module is restricted to using only (single-layer) S-R learning, e.g., the LMS rule of Widrow & Hoff (1960). Hippocampal-lesion experiments are modeled by removing the hippocampal module. This results in the bottom layer of the cortical module remaining fixed (e.g., with non-modifiable weights). The upper layer of the cortical module, however, can still be trained to learn based on the fixed recoding of the cortical inputs which occurs in the cortical bottom layer. Learning without the hippocampal module is thus limited to discriminations which can be solved without learning a new stimulus representation.*

hidden layer is provided as feedback to the cortical module hidden layer, the cortical module can use its simple learning rules to map from sensory inputs to this representation and from the representation to outputs, allowing it to learn complex associations. This is our connectionist conceptualization of normal, intact cortico-hippocampal interaction in associative learning.

Within this framework, a hippocampal lesion is characterized by removing the hippocampal module. This has the effect of eliminating the hippocampal feedback which the cortical module would otherwise use to construct hidden layer representations. When the hippocampal-system module is so lesioned, the lower layer of cortical module weights remains fixed, and the system learns associative stimulus-response (S-R) relationships based on this fixed encoding (representation) of the stimulus inputs. In comparison, the intact system learns the same S-R relationships, but does so based on a flexible and distributed re-coding of the stimulus inputs which reflect both predictive S-R relationships as well as sensory-sensory correlations. For example, simple discrimination learning (A+/B-) is largely unaffected or even facilitated after hippocampal lesion (Schmaltz & Theios, 1972; Eichenbaum, et al., 1988). As shown in Figure 2A, both the intact and the lesioned network models can solve a simple discrimination task. Furthermore, the lesioned network shows some facilitation. This occurs because the initial representation is sufficient to learn the task, whereas in the intact model, this initially sufficient representation is altered by hippocampal influence, retarding learning.

Applications to Data

We turn now to examining the behavior of the cortico-hippocampal model in several other key paradigms. These simulations will illustrate how the model instantiates or refines aspects of four prior

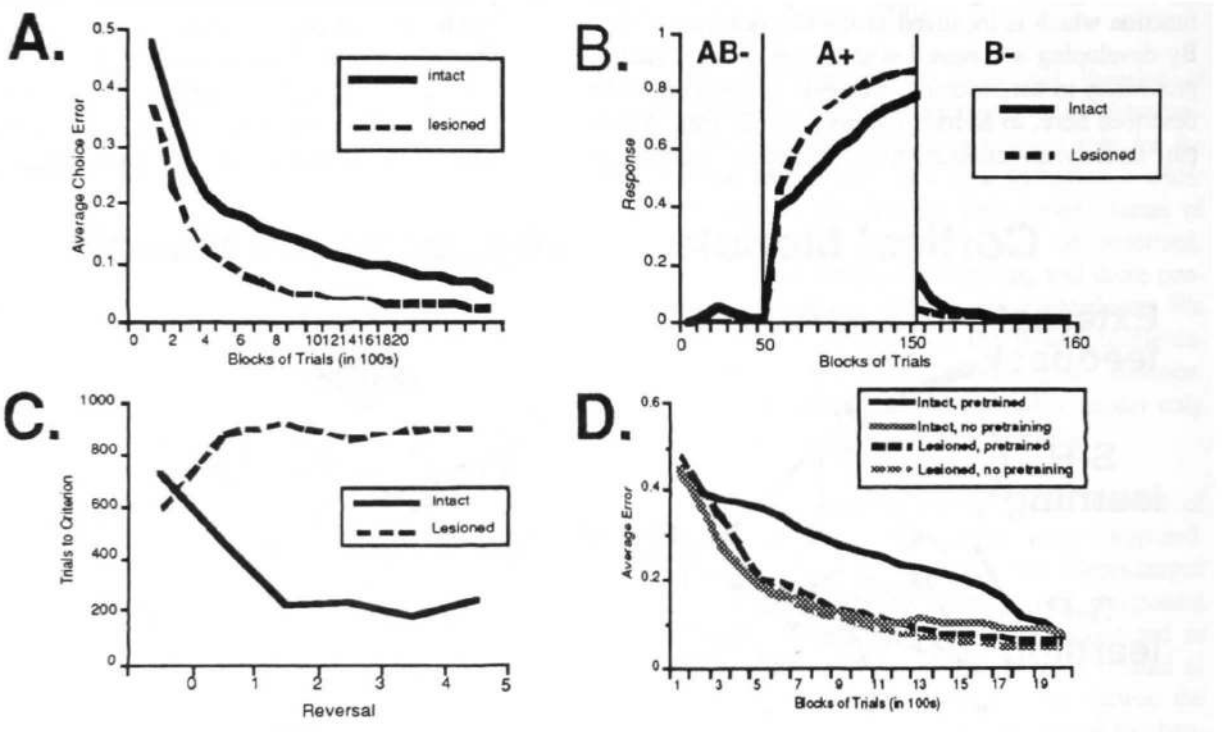

Figure 2. *Simulations of intact and lesioned cortico-hippocampal model.* **(A) Simple Discrimination**: *Training to A+/B-. Lesioned model learns somewhat faster as intact model must first learn a new representation in the hippocampal module and then transfer this representation to cortical module.* **(B) Sensory Preconditioning**: *Pre-exposure to a non-reinforced AB- compound followed by training to A+. Intact and lesioned systems are similar through first two phases, but only intact system shows transfer of response when tested with B in third phase.* **(C) Multiple Reversals**: *Training on A+/B-, followed by reversal A-/B+, then A+/B-, etc. Intact system shows a progressive decrease in the number of trials required to learn each discrimination; lesioned system has difficulty with all but the first discrimination.* **(D) Latent Inhibition**: *Non-reinforced A- training followed by A+ training. Intact but not lesioned system impaired on A+ learning compared with control condition of pre-training to another cue (e.g., C+).*

information-processing theories of hippocampal-system function: chunking, stimulus-selection, cue-configuration and contextual labeling.

Chunking

As suggested by Wickelgren (1979), the model incorporates a "chunking" mechanism through the sensory compression which occurs in the hippocampal-system module to static and co-occurring (most clearly redundant) features, including the context. This is seen in **sensory preconditioning**, where an animal is first pre-exposed to an unreinforced AB stimulus compound, and then given A+ training. In a final training phase, the animal shows partial transfer of the learned response to stimulus cue B. Port & Patterson (1984) demonstrated that the hippocampal-region is necessary for sensory preconditioning. Figure 2B shows that our model is consistent with this result. The intact system builds an internal recoding during pre-exposure to AB- training which "chunks" A and B together. This chunked representation persists during the A+ training. Therefore, some of the A+ association transfers to B through a process akin to an acquired form of stimulus generalization.

Stimulus Selection

Theories of stimulus selection in classical conditioning can be differentiated into two classes: those based on a modulation of the reinforcing value of the unconditioned stimulus (e.g., Rescorla & Wagner, 1972) and those which presume an attentional or salience modulation of sensory inputs (e.g., Pearce & Hall, 1980; Mackintosh, 1975). Our hippocampal-system model incorporates both forms of stimulus selection. Reinforcement modulation is localized in the cerebellar and cerebral cortices, while sensory modulation localized in the hippocampal region. This mapping is consistent with results indicating that behaviors which are uniquely explained by sensory modulation (e.g., reversal facilitation, latent inhibition) show the clearest deficits after hippocampal lesion. In comparison, stimulus selection behaviors which can be uniquely explained by reinforcement modulation (e.g., conditioned inhibition) show no hippocampal deficit. Phenomena which can be explained by both mechanisms (and hence are assumed in our theory to be multiply determined across several brain regions) have resulted in inconclusive or partial deficits (e.g., blocking and overshadowing). We focus now on the two phenomena which we expect to show clearest hippocampal dependence.

In reversal learning, an animal is first trained on a simple A+/B- discrimination. This is then followed by reversal training on A-/B+. These two discriminations are then repeatedly reversed. Normal intact animals shown a progressive facilitation in learning the new discriminations; in contrast, HL animals show an impairment (Berger & Orr, 1983). As illustrated in Figure 2C, the intact cortico-hippocampal system shows a progressive decrease in the number of trials required to learn each reversal – reflecting the fact that the hippocampal-module's distributed stimulus recoding devotes more and more resources to the relevant cues. The lesioned system, with no such mechanism for stimulus selection, must first "unlearn" previous discriminations before starting afresh to learn each new reversal.

Latent inhibition, first described by Lubow (1973), is an especially important "marker" of hippocampal processing because the hippocampal-damaged animals show *increased* responding in a transfer task. When animals are first given A- (unreinforced) trials, and then switched to A+ (reinforced) training, their acquisition of A+ is impaired relative to animals with no A- pre-training. Error-correcting models such as the Rescorla-Wagner (1972) model and error backpropagation networks (Rumelhart, et al. 1986) fail to predict latent inhibition, because they expect no learning during the (errorless) A- training phase. In our model, however, sensory-sensory learning does take place during A- pre-training. The hippocampal module learns to "chunk" A together with the context because neither is predictive of the US or of any other significant event. Later, in A+ training, the representation of A must be "de-chunked" from the context before learning can occur (Figure 2D). Solomon and Moore (1975) showed that latent inhibition is absent in HL animals, and Figure 2D shows that it is absent in the lesioned system as well.

Cue-Configuration

Sutherland and Rudy's (1989) cue-configuration theory proposes that the hippocampus is necessary for the acquisition and retention of configural associations. Our cortico-hippocampal model implies a similar hippocampal involvement in configural learning: configural tasks will typically entail a stimulus recoding necessitating an intact hippocampus. Simulation results (not shown here) demonstrate a lesioned system deficit for configural tasks such as negative pat-

terning (A+/B+/AB-).

Contextual Labeling

Our model can also be viewed as an instantiation of theories suggesting a key role for the hippocampus in developing a "contextual tag" for stimulus-response associations (Hirsh, 1974; Nadel & Willner, 1989, Winocur, et al., 1987). Hippocampal-lesioned animals are often shown to have difficulty encoding context. Given training with A+ in one context, normal intact animals show a decreased response when tested with A in a new context; HL animals show no such decrease (Penick & Solomon, 1991). Likewise (simulations not shown), the intact cortico-hippocampal model shows a decreased response if contextual cues are changed; the lesioned model shows no such deficit. In the intact model, the contextual cues which co-occur with A become part of that stimulus's representation; when A is shown in a different context, the representation is less strongly activated than usual, and the response strength drops.

Summary and Discussion

The model we have presented here shows how a specific network architecture can form compressed, predictive, and distributed representations of stimuli, which are made available to other learning systems (such as the cerebellum and cerebral cortex). This model incorporates and refines aspects of many prior, qualitative information-processing theories of hippocampal function, including chunking (Wickelgren, 1979), stimulus selection Rescorla & Wagner, 1972; Pearce & Hall, 1980; Mackintosh, 1975), contextual labeling (Hirsh, 1974; Nadel & Willner, 1989) and cue-configuration (Sutherland & Rudy, 1989). Our theory also relates to other behaviors sensitive to hippocampal damage. For example, several task-specific theories of hippocampal-region function have noted the impairment of HL animals in spatial navigation (O'Keefe & Nadel, 1978). In our connectionist cortico-hippocampal model, place learning could be another kind of representational learning; the hippocampus would be responsible for mapping from a partial view of an environment into a full representation of a place. Linear autoassociator models of the hippocampus (e.g., McNaughton, 1989; Rolls, 1990) have previously been proposed. A predictive autoencoder, used here as a conceptualization of the hippocampal-system, generalizes the properties of a linear autoassociator.

In its current form, the model does not address hippocampal mediation of temporal and sequential processing, functional roles implied by the failure of hippocampal-damaged animals at conditioning with long ISI delays or trace conditioning (Moyer, et al. 1990). These additional temporal roles may either be interpreted as requiring refinements of the same mechanisms proposed here, or they may be localized within different brains structures in the medial temporal lobe. Future efforts will be needed to better understand the interaction between temporal and representational processing in the hippocampal-region and the precise neurobiological locus (or loci) and physiological characteristics of these two functions.

Acknowledgments

This work was supported by a Young Investigator Program award from the Office of Naval Research to MAG. For their helpful suggestions and critical feedback, we are indebted to Gyorgy Buzsaki, Ian Creese, Howard Eichenbaum, Paul Glauthier, Bruce McNaughton, Lynn Nadel, Jerry Rudy, Larry Squire, and Richard Thompson.

References

Berger, T., and Orr, W. 1983. Hippocampectomy disrupts discrimination reversal learning of the rabbit nictitating membrane response. *Behavioral Brain Research* 8:49-68.

Buzsaki, G., 1989. Polysynaptic long-term potentiation: A physiological role of the perforant path CA3/CA1 pyramidal cell synapse. *Brain Research* 45:192-195.

Cohen, N. and Squire, L., 1980. Preserved learning and retention of pattern analysing skill in amnesia: Dissociation of knowing how and knowing that. *Science* 210:207-209.

Douglas, R. and Pribram, K., 1966. Learning and limbic lesions. *Neuropsychologia* 4:192-220.

Eichenbaum, H. and Buckingham, J., 1991. Studies on hippocampal processing: Experiment, theory and model. In *Memory: Organization and locus of change*, ed. J. McGaugh, Oxford University Press, Oxford (in press).

Eichenbaum, H.; Fagan, A.; Mathews, P.; and Cohen, N., 1988. Hippocampal system dysfunction and odor discrimination learning in rats: Impairment or facilitation depending on representational demands. *Behavioral Neuroscience* 102:331-339.

Gluck, M. and Bower, G., 1988. Evaluating an adaptive network model of human learning. *Journal of Memory and Language* 27:166-195.

Grastyan, E.; Lissak, K.; Madarasz, I.; Donhoffer, H., 1959. Hippocampal electrical activity during the development of conditioned reflexes. *Electroencephalography and Clinical Neurophysiology* 11:409-430.

Hinton, G., 1989. Connectionist learning procedures. *Artificial Intelligence* 40:185-234.

Hirsh, R., 1974. The hippocampus, conditional operations and cognition. *Physiological Psychology* 8:175-182.

Lubow, R., 1973. Latent inhibition. *Psychological Bulletin* 79:398-407.

Lynch, G., and Granger, R., 1991. Serial steps in memory processing: Possible clues from studies of plasticity in the olfactory-hippocampal circuit. In, *Olfaction as a model system for computational neuroscience*, ed. J. Davis, MIT Press, Cambridge MA.

Mackintosh, N., 1975. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review* 82:276-298.

McNaughton, B., 1989. Neuronal mechanisms for spatial computation and information storage. In, *Neural connections, mental computations*, ed. R. Harnish, MIT Press, Cambridge MA.

McNaughton, B.; Chen, L.; and Markus, E., 1991. "Dead reckoning", landmark learning and the sense of direction: A neurophysiological and computational hypothesis. *Journal of Cognitive Neuroscience* 3:190-202.

McNaughton, B. and Nadel, L., 1990. Hebb-Marr networks and the neurobiological representation of action in space. In, *Neuroscience and Connectionist Theory*, ed. D. Rumelhart, Lawrence Erlbaum Associates, Hillsdale NJ.

Mishkin, M., 1982. Memory in monkeys severely impaired by combined but not separate removal of the amygdala and hippocampus. *Nature* 273:297-298.

Morris, R.; Garrud, P.; Rawlins, J.; and O'Keefe, J., 1982. Place navigation impaired in rats with hippocampal lesions. *Nature* 297:681-683.

Moyer, J.; Deyo, R.; and Disterhoff, J., 1990. Hippocampectomy disrupts trace eye-blink conditioning in rabbits. *Behavioral Neuroscience* 104:243-252.

Nadel, L. and Willner, J., 1989. Context and conditioning: A place for space. *Physiological Psychology* 8:218-228.

O'Keefe, J. and Nadel, L., 1978. *The Hippocampus as a Cognitive Map*, Claredon University Press, Oxford UK.

O'Keefe, J.; Nadel, L.; Keightly, S.; and Kill, D., 1975. Fornix lesions selectively abolish place learning in the rat. *Experimental Neurology* 48:152-166.

Pearce, J. and Hall, G., 1980. A model for Pavlovian learning: Variations in the effectiveness of conditioned by not of unconditioned stimuli. *Psychological Review* 87:532-552.

Penick, S. and Solomon, P., 1991. Hippocampus, context and conditioning. *Behavioral Neuroscience* 105:611-617.

Port, R. and Patterson, M., 1984. Fimbrial lesions and sensory preconditioning. *Behavioral Neuroscience* 98:584-589.

Rescorla, R. and Wagner, A., 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In, *Classical Conditioning II: Current Research and Theory*, ed. W. Prokasy, New York.

Rolls, E., 1990. Theoretical and neurophysiological analysis of the functions of the primate hippocampus in memory. *Cold Spring Harbor Symposia on Quantitative Biology* 55:995-1006.

Rumelhart, D.; Hinton, G.; Williams, R., 1986. Learning internal representations by error propagation. In, *Parallel Distributed Processing - Explorations in the Microstructure of Cognition*, vol. 1, eds. D. Rumelhart & J. McClelland, MIT Press, Cambridge MA.

Schmajuk, N. and Moore, J., 1985. Real-time attentional models for classical conditioning and the hippocampus. *Physiological Psychology* 13:278-290.

Schmaltz, L. and Theios, J., 1972. Acquisition and extinction of a classically conditioned response in hippocampectomized rabbits (Oryctolagus cuniculus). *Journal of Comparative and Physiological Psychology* 79:328-333.

Scoville, W. and Millner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* 20:11-21.

Solomon, P. and Moore, J., 1975. Latent inhibition and stimulus generalization of the classically conditioned nictitating membrane response in rabbits (Oryctolagus cuniculus) following dorsal hippocampal ablation. *Journal of Comparative and Physiological Psychology* 202:1192-1203.

Squire, L., 1987. *Memory and Brain*, Oxford University Press, New York.

Squire, L. and Zola-Morgan, S., 1983. The neurology of memory: Qualitative assessment of retrograde amnesia in two groups of amnesic patients. *Journal of Neuroscience* 9:828-839.

Sutherland, R. and Rudy, J., 1989. Configural association theory: The role of the hippocampal formation in learning, memory and amnesia. *Psychobiology* 17:129-144.

Wickelgren, W., 1979. Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review* 86:44-60.

Winocur, G.; Rawlins, J.; and Gray, J., 1987. The hippocampus and conditioning to contextual cues. *Behavioral Neuroscience* 101:617-625.