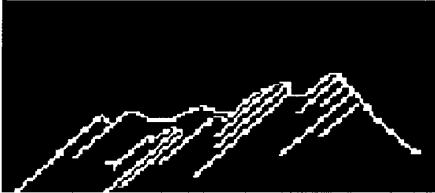


Institute of Cognitive Science



# Technical Report

University of Colorado, Boulder

## Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary Learning Systems Approach

Kenneth A. Norman and Randall C. O'Reilly  
norman@psych.colorado.edu oreilly@psych.colorado.edu

Department of Psychology  
University of Colorado  
Boulder, CO 80309

Technical Report 01-02 – Version 10th August, 2001

## Abstract

We present a computational neural network model of recognition memory based on the biological structures of the hippocampus and medial temporal lobe cortex (MTLC), which perform complementary learning functions. The hippocampal component of the model contributes to recognition by recalling specific studied details. MTLC can not support recall, but it is possible to extract a scalar familiarity signal from MTLC that tracks how well the test item matches studied items. We present simulations that establish key qualitative differences in the operating characteristics of the hippocampal recall and MTLC familiarity signals, and we identify several manipulations (e.g., target-lure similarity, interference) that differentially affect the two signals. We also use the model to address the stochastic relationship between recall and familiarity (i.e., are they independent), and the effects of partial vs. complete hippocampal lesions on recognition.

Contents	
<b>Introduction</b>	<b>3</b>
Mathematical Models of Recognition Memory .	4
Cognitive Neuroscience Approaches to Recognition Memory . . . . .	5
Summary: Combining the Approaches . . . . .	6
<b>The Complementary Learning Systems Model</b>	<b>6</b>
The Cortical Model . . . . .	8
Familiarity as Sharpening . . . . .	8
The Hippocampal Model . . . . .	9
The Hippocampal Recall Measure . . . . .	10
Simulation Methodology . . . . .	11
<b>Part I: Basic Network Properties</b>	<b>12</b>
<b>Simulation 1: Pattern Separation and Blending</b>	<b>12</b>
<b>Simulation 2: ROC Curves</b>	<b>13</b>
<b>Simulation 3: Variability and Scaling Effects</b>	<b>16</b>
Sampling Variability . . . . .	16
Other Sources of Variability . . . . .	17
How Variability is Implemented in the Models .	18
<b>Part II: Applications To Behavioral Phenomena</b>	<b>18</b>
<b>Simulation 4: Lure Relatedness and Test Format</b>	<b>18</b>
Interactions	18
YN Performance . . . . .	18
FC Performance . . . . .	20
The Cortical Model . . . . .	20
Simulations With Encoding Variability .	21
The Hippocampal Model . . . . .	21
Testing the Model's Predictions . . . . .	22
<b>Simulation 5: Associative Recognition and Sensi- tivity to Conjunctions</b>	<b>23</b>
Effects of Test Format . . . . .	24
Tests of the Model's Predictions . . . . .	25
<b>Simulation 6: Interference and List Strength</b>	<b>26</b>
General Principles of Interference in our Models	27
List Strength Results . . . . .	28
Interference in the Hippocampal Model . . . . .	28
Interference in the Cortical Model . . . . .	30
Boundary Conditions on the Null LSE . .	31
Differentiation . . . . .	31
Summary and Predictions . . . . .	32
Testing the Model's Predictions . . . . .	32
Confidence Ratings . . . . .	33
Self-Report Measures . . . . .	33
Lure Relatedness . . . . .	34
Summary . . . . .	34
<b>Simulation 7: List Length and Dissociations with List Strength</b>	<b>35</b>
Decay and Delay Predictions . . . . .	36
<b>Simulation 8: The Combined Model and the In- dependence Assumption</b>	<b>37</b>
Interference Induced Decorrelation . . . . .	37
Effects of Other Kinds of Variability . . . . .	39
Summary and Implications . . . . .	39
<b>Simulation 9: Lesion Effects in the Combined Model</b>	<b>40</b>
Lesion Controversies . . . . .	40
Partial Lesion Simulations . . . . .	40
Region-Specific Hippocampal Lesions . . . . .	42
MTLC Lesions . . . . .	43
Summary and Implications . . . . .	43
<b>General Discussion</b>	<b>44</b>
Summary of Key Simulation Results . . . . .	44
Comparison with other Theories and Models . .	46
Aggleton & Brown's Recall/Familiarity Theory . . . . .	46
Yonelinas & Jacoby's Dual-Process Sig- nal Detection Model . . . . .	46
Bayesian Global Matching Models (e.g., REM) . . . . .	47
Ratcliff's (1990) Neural Network Recognition Model . . . . .	48
Other Neural Network Models of Hip- pocampus and Cortex . . . . .	48
Alternate Dependent Measures . . . . .	50
Future Directions . . . . .	51
Conclusion . . . . .	51
<b>Acknowledgements</b>	<b>51</b>
<b>Appendix A: The Leabra Algorithm</b>	<b>51</b>
Pseudocode . . . . .	52
Point Neuron Activation Function . . . . .	52
k-Winners-Take-All Inhibition . . . . .	52
Hebbian Learning . . . . .	53
Weight Contrast Enhancement . . . . .	53
<b>Appendix B: Basic Parameters</b>	<b>53</b>
<b>Appendix C: Fast Weight Mechanisms</b>	<b>54</b>
<b>References</b>	<b>55</b>

## Introduction

Memory can be subdivided according to functional categories (e.g., declarative vs. procedural memory; Squire, 1992; Cohen & Eichenbaum, 1993), and according to neural structures (e.g., hippocampally-dependent vs. non-hippocampally-dependent forms of memory). Various attempts have been made to align these functional and neural levels of analysis, e.g., Squire (1992) and others have argued that declarative memory depends on the medial temporal lobe, whereas procedural memory depends on other cortical and subcortical structures. Recently, we and our colleagues have set forth a computationally explicit theory of how hippocampus and neocortex contribute to learning and memory (the *Complementary Learning Systems* model; O'Reilly & Rudy, 2001; McClelland, McNaughton, & O'Reilly, 1995). In this paper, we advance the Complementary Learning Systems model by using it to provide a comprehensive treatment of recognition memory performance.

Recognition memory refers to the process of identifying stimuli or situations as having been experienced before, for example when you recognize a person you know in a crowd of strangers. Recognition can be compared with various forms of recall memory where specific content information is retrieved from memory and produced as a response; recognition does not require recall of specific details (e.g., one can recognize a person as being familiar without being able to recall who exactly they are or where you know them from). Nevertheless, recognition can certainly benefit from recall of specific information — if you can recall that the person you recognized at the supermarket is in fact your veterinarian, that reinforces your feeling that you actually do know this person.

The fact that recognition memory can be subserved by these two qualitatively different types of memory signals (non-specific familiarity and specific recall; Mandler, 1980; Jacoby, Yonelinas, & Jennings, 1997) makes it a particularly rich and interesting domain for understanding the different contributions of underlying neural systems — is it possible that different brain areas subserve these qualitatively distinct functions? Recently, Aggleton & Brown have suggested, based on a variety of empirical findings, that the hippocampus is critically important for recall, while surrounding medial temporal lobe cortical areas (in particular, the perirhinal cortex) can provide a non-specific familiarity signal (Aggleton & Brown, 1999; Brown & Aggleton, 2001). Our model incorporates this general division of labor, but provides for considerably more precise and subtle characterizations of the differential contributions of these brain areas, based on the functional characteristics of their respective biological substrates.

Insofar as our model provides a computationally explicit account of recognition memory, and is capable of simulating human performance on a variety of recognition memory tests, it can be compared to other mathematical memory models (e.g., Shiffrin & Steyvers, 1997; McClelland & Chappell, 1998; Hintzman, 1988; Gillund & Shiffrin, 1984). The most salient difference between our model and other mathematical models is that other models are *abstract* — they describe the memory system in broad functional terms that do not relate specifically to underlying brain systems. Because our model makes specific claims about the brain basis of recognition, the research presented here also ties in to the broader cognitive neuroscience literature that characterizes functional properties of the hippocampus and cortical areas on the basis of data from anatomy, physiology, the effects of lesions, and various neural recording/imaging techniques (e.g., Squire & Zola, 1996; Schacter, Wagner, & Buckner, 2000; Nyberg & Cabeza, 2000). Therefore, we view our model as providing a critical bridge between mathematical modeling and cognitive neuroscience approaches to recognition memory, which have to this point been pursued in relative isolation from each other. By making this bridge, we can take advantage of constraints from neuroscience to inform computational models of recognition memory, and (reciprocally) we can use computational modeling to inform the debate over how different brain structures contribute to recognition.

The remainder of the paper is organized as follows. In the next two sections, we will describe two questions that have proved challenging for math modeling and cognitive neuroscience approaches to recognition, respectively: In the math modeling literature, there has been considerable controversy regarding how to characterize the contribution of recall (vs. familiarity) to recognition memory; in the cognitive neuroscience literature, researchers have debated how the hippocampus (vs. medial temporal lobe cortex) contributes to recognition. An important goal of the paper is to show how our model provides specific answers to these puzzles. Then in the main body of the paper, we present the *Complementary Learning Systems* (CLS) neural network model of recognition memory; the CLS model consists of two components: a hippocampal network, and a cortical network. We show how the hippocampal network provides a recall signal that discriminates between studied and nonstudied items; likewise, we show how the cortical network provides a *familiarity* signal that also discriminates between studied and nonstudied items.

We present simulations characterizing the basic properties of these signals, focusing on differences in how information is represented in the two networks, and how these representational differences result in the two sig-

nals having different operating characteristics. We then build on these basic simulation results by applying the model to specific recognition memory paradigms; in this context, we document key qualitative differences in how lure relatedness and interference manipulations affect the two signals. In our final set of simulations, we look at how the two systems interact with one another; using a combined cortico-hippocampal network, we explore how different factors affect the statistical relationship between recall and familiarity, and we explore how partial and complete hippocampal and cortical lesions affect overall recognition performance. We conclude by discussing the relationship between this work and other theoretical perspectives.

### *Mathematical Models of Recognition Memory*

The central feature of existing math models of recognition memory is that they try to explain recognition performance in terms of a unitary familiarity process that indexes — in a holistic fashion — how well the test probe *matches* all of the items stored in memory. Math modelers have focused on explaining behavioral data from list-learning experiments; within this domain, these single-process models have been able to account for an impressively wide range of findings (for reviews, see Clark & Gronlund, 1996; Raaijmakers & Shiffrin, 1992; Ratcliff & McKoon, 2000). This single-process approach can be contrasted with the dual-process approach that we and others have taken, which posits that subjects can make recognition decisions based on a holistic familiarity process or based on recall of specific studied details.

The potential for recall to make a distinct contribution to recognition is most clearly illustrated by paradigms where recall and familiarity are placed in opposition (Jacoby, 1991) — i.e., where subjects can use recall to reject lures that are familiar (either because they were presented outside of the study list, or because they resemble studied items). For example, Hintzman, Curran, and Oppy (1992) had participants study singular and plural words; a given word was studied in its singular or plural form (but never both). At test, participants had to discriminate between studied words, switched-plurality lures (e.g., study “RATS”, test “RAT”), and unrelated lures (i.e., lures where the word was not studied in singular or plural form). Switched-plurality lures will be familiar, but they can be rejected if subjects recall studying that item in a different plurality (e.g., if they remember “I didn’t study RAT, I studied RATS”). Hintzman et al. (1992) found that false recognition of switched-plurality lures first increased, then (in some experiments) decreased as a function of how many times the corresponding studied item was presented; they argued that repeating an item at study made the corresponding lure

more familiar (accounting for the initial increase in false recognition), but also increased the odds that the corresponding lure would trigger recall of the studied plurality (accounting for the subsequent decrease in false recognition).

There has been extensive debate regarding whether subjects actually utilize recall to reject similar lures in this paradigm (e.g., Rotello, 2000). However, it is clear that similar lures do sometimes trigger recall of the corresponding studied item, and that subjects could (if they so chose) use this recalled information to reject switched-plurality lures. Also, it should be clear that recall can be used as evidence that a test item was studied (when recalled information matches the test probe).

The two main reasons why math models do not routinely incorporate a recall process are parsimony, and lack of behavioral constraints. As mentioned above, familiarity-only models can explain a wide range of recognition findings — even findings, that at first glance, appear to require a recall process; for example, McClelland and Chappell (1998) showed how a single-process model can account for the Hintzman et al. (1992) data described above. There are very few findings that uncontroversially necessitate a recall process, and many of these findings come from specialized paradigms like Jacoby’s process dissociation procedure (Jacoby, 1991; see Ratcliff, Van Zandt, & McKoon, 1995 for discussion of how single-process models can account for process-dissociation data). As such, it is always possible to treat these findings as special cases that have little relevance to performance on standard item-recognition tests. If it is possible to account for the vast majority of the recognition dataspace using a single-process model, there would be no point in positing a recall process. However, even if a modeler wanted to build a dual-process model with distinct recall and familiarity processes, they would run into the second problem — lack of reliable behavioral constraints.

To constrain a dual-process model using behavioral data, one needs some way of measuring the separate contributions of recall and familiarity. Over the last decade, Jacoby, Yonelinas, and colleagues have devised several techniques for quantitatively measuring the contributions of recall and familiarity to recognition performance (ROC analysis: Yonelinas, Dobbins, Szymanski, & Dhaliwal, 1996; independence remember-know: Jacoby et al., 1997; process dissociation: Jacoby, 1991; Yonelinas & Jacoby, 1996; see also Yonelinas, *in press*). All of these techniques rely on a core set of assumptions about recall and familiarity: For example, they assume that recall and familiarity are stochastically independent, and that recall is a *high-threshold* process (meaning that recall is all-or-none, and that it never occurs for lure

items). These assumptions are very controversial, especially the independence assumption (Curran & Hintzman, 1995). Testing the assumptions brings up a number of chicken-and-egg problems. For example, one needs to measure familiarity to assess its independence from recall, but one needs to assume independence to measure familiarity. These chicken-and-egg problems have led to a rift between math modelers and other memory researchers. On the empirical side, there is now a vast body of data on recall and familiarity, gathered using these assumption-laden measurement techniques — this data could potentially be used to constrain dual-process models. However, on the theoretical side, modelers are not making use of this data because of reasonable concerns about the validity of the assumptions used to collect this data. To resolve this impasse, we need some other source of evidence that we can use to constrain dual-process models. We show that if one pays attention to how the brain implements recall and familiarity, this can provide a critical source of constraints for dual-process models.

### *Cognitive Neuroscience Approaches to Recognition Memory*

Much of what we know about the brain basis of recognition comes from the study of medial temporal lobe amnesics — these patients typically have lesions encompassing both the hippocampus and the *medial temporal lobe cortices* (MTLC, including perirhinal, entorhinal, and parahippocampal cortex) surrounding the hippocampus.

Patients with medial temporal damage show impaired recall and recognition but intact performance on other memory tests (e.g., perceptual priming, skill learning). One possible explanation for why recall and recognition depend on the medial temporal region is that both tasks require participants to form associations between “core” item attributes (i.e., aspects of an item’s representation that vary minimally from context to context) and contextually-varying attributes (e.g., the font that a word is presented in; its position on the screen; elaborations elicited by the encoding task) — recognition judgments as studied in the laboratory do not just ask, did you *ever* see this item; rather, they ask, did you see this item *in a particular context* (presented in a particular manner, at a particular time and location). The medial temporal region is the only part of the brain that is set up to associate widely different types of information — it is located on top of the cortical hierarchy and therefore is ideally positioned to associate aspects of the current episode that are being processed in domain-specific cortical modules (e.g., Mishkin, Suzuki, Gadian, & Vargha-Khadem, 1997; Mishkin, Vargha-Khadem, & Gadian, 1998) (Figure 1).

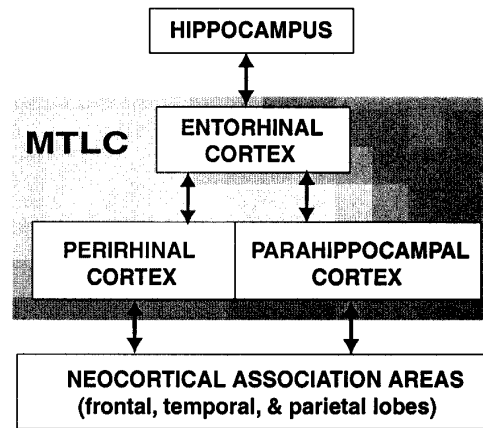


Figure 1: Schematic box diagram of neocortex, MTLC, and hippocampus. Medial temporal lobe cortex serves as the interface between neocortex and the hippocampus. Medial temporal lobe cortex is located at the very top of the cortical processing hierarchy — it receives highly processed outputs of domain-specific cortical modules, integrates these outputs, and passes them on to the hippocampus; it also receives output from the hippocampus, and passes this activation back to domain-specific cortical modules via feedback connections.

The finding of impaired recall and recognition in medial temporal amnesics is the basis for several influential taxonomies of memory. Most prominently, Squire, Cohen, and others have argued that the medial temporal lobes implement a *declarative* memory system, which supports recall and recognition, and that other brain structures support *procedural* memory (e.g., perceptual priming, motor skill learning; Squire, 1992, 1987; Cohen & Eichenbaum, 1993; Cohen, Poldrack, & Eichenbaum, 1997; Eichenbaum, 2000). Recently, researchers have sought to tease apart the contributions of different medial temporal structures to declarative memory by looking at more focal lesions, especially focal hippocampal damage (sparing MTLC). If the hippocampus is involved in both recognition and recall, then both recall and recognition deficits should be present after focal hippocampal damage. Consistent with this prediction, these patients show severely impaired recall. The surprising finding is that recognition is sometimes intact after focal hippocampal damage in humans (Vargha-Khadem, Gadian, Watkins, Connelly, Van Paesschen, & Mishkin, 1997; Holdstock, Mayes, Roberts, Cezayirli, Isaac, O'Reilly, & Norman, in press; Mayes, Holdstock, Isaac, Hunkin, & Roberts, submitted; but see Reed & Squire, 1997; Manns & Squire, 1999; Rempel-Clower, Zola, & Amaral, 1996; Zola-Morgan, Squire, & Amaral, 1986; Reed, Hamann, Stefanacci, & Squire, 1997,

all of which found impaired recognition after focal hippocampal lesions). The monkey literature parallels the human literature — some studies have found relatively intact recognition (indexed using the delayed nonmatch-to-sample test) following focal hippocampal damage (e.g., Murray & Mishkin, 1998) whereas others have found impaired recognition (e.g., Zola, Squire, Teng, Sefanacci, Buffalo, & Clark, 2000; Beason-Held, Rosene, Killiany, & Moss, 1999). Spared recognition following hippocampal lesions depends critically on MTLC — whereas recognition is sometimes spared by focal hippocampal lesions, it is never spared after lesions that encompass both MTLC and the hippocampus (e.g., Aggleton & Shaw, 1996).

This data can be summarized from the dual-process perspective (e.g., Aggleton & Brown, 1999):

- The hippocampus supports recall.
- The MTLC can support some degree of (familiarity-based) recognition on its own.

This framework captures, at a gross level, how hippocampal damage affects memory, but it is too vague to be useful in explaining *variability* across tests and patients in how hippocampal damage affects recognition. According to this framework, recognition impairments — when they occur — are due to the loss of the hippocampal recall process. However, in the absence of further specification of this hippocampal contribution (and how it differs from the contribution of MTLC), it is not possible to proactively determine whether this contribution will be missed on a given test.

To explain spared item recognition performance after focal hippocampal damage (e.g., Vargha-Khadem et al., 1997), Aggleton and Brown (1999) argue that the hippocampus is required to form new associations, but cortex can support memory for individual items or features on its own. This implies that item memory should be intact but associative memory should be impaired after focal hippocampal damage. However, Andrew Mayes and colleagues have found that hippocampally-lesioned patient YR, who shows intact performance on some item recognition tests (e.g., Mayes et al., submitted), shows impaired performance on other item recognition tests (e.g., Holdstock et al., in press), and spared performance on some associative recognition tests (which require subjects to associate unrelated stimuli, e.g., the words “window” and “reason”; Mayes, Isaac, Downes, Holdstock, Hunkin, Montaldi, MacDonald, Cezayirli, & Roberts, 2001). Thus, it is becoming increasingly evident that the effects of hippocampal damage are complex — it is unlikely that simple dichotomies (like item vs. associative memory) will be sufficient to describe the respective contributions of hippocampus and MTLC to recognition.

### Summary: Combining the Approaches

What should be clear at this point is that the math modeling and cognitive neuroscience approaches to recognition memory would greatly benefit from increased crosstalk: Math modeling approaches need a new source of constraints before they can fully explore how recall contributes to recognition; and cognitive neuroscience approaches need a new, more mechanistically sophisticated vocabulary for talking about the roles of different brain structures in order to adequately characterize differences in how MTLC vs. hippocampus contribute to recognition.

The goal of our research is to achieve a synthesis of these two approaches, by constructing a computational model of recognition memory in which there is a transparent mapping between different parts of the model and different subregions of hippocampus and MTLC. This mapping makes it possible to address neuroscientific findings using the model. For example, to predict the effects of a particular kind of hippocampal lesion, we can “lesion” the corresponding region of the model. By bringing a wide range of constraints — both purely behavioral and neuroscientific — to bear on a common set of mechanisms, we hope to achieve a more detailed understanding of how recognition memory works.

### The Complementary Learning Systems Model

Our overall view of cortical and hippocampal processing builds on the Complementary Learning Systems (CLS) framework (O'Reilly & Rudy, 2001; McClelland et al., 1995). The central idea of this framework is that there are two kinds of learning that people and other animals need to do to successfully negotiate the environment:

- Memorize specific events.
- Extract the general structure of the environment (e.g., statistically regular patterns of co-occurrence).

Furthermore, this framework posits that learning about specifics and extracting generalities are computationally incompatible tasks, so we have evolved specialized neural systems for performing these tasks (Figure 2; for a contrasting perspective, see Carpenter & Grossberg, 1993).

If you want to remember specific events (e.g., where you parked your car today), you need to block out interference from where you parked yesterday, and the day before, and so on. The best way to minimize interference is to assign *separate representations* to events, no

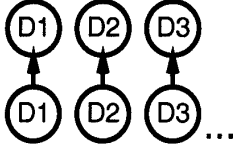
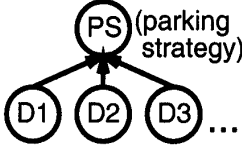
Two Incompatible Goals		
	Remember Specifics	Extract Generalities
Example:	Where is car parked?	Best parking strategy?
Need to:	Avoid interference	Accumulate experience
<i>Solution:</i>		
1.	Separate reps (keep days separate) 	Overlapping reps (integrate over days) 
2.	Large learning rate (encode immediately)	Small learning rate (integrate over days)
<i>These are incompatible, need two different systems:</i>		
System:	Hippocampus	Neocortex

Figure 2: Computational motivation for two complementary learning & memory systems in the brain: There are two incompatible goals that such systems need to solve. One goal is to remember specific information, in this example where one's car is parked on a specific day. The other goal is to extract generalities across many experiences, for example in determining the best overall parking strategy. The neural solutions to these goals are incompatible: Memorizing specifics requires separate representations that are learned quickly, while extracting generalities requires overlapping representations (to represent commonalities across events) and a small learning rate (such that no single event dominates the representation). Thus, it makes sense to have two separate neural systems that are optimized for each of these goals.

matter how similar they are; using separate representations helps keep all of your different parking memories from blending together. Also, to memorize specific (possibly transient) events, it is necessary to learn quickly by using a *large learning rate* (because you only have one chance to memorize the event before it goes away). In contrast, if you want to determine the best overall parking strategy, you need to accumulate experience — you need to integrate across parking episodes to get a statistical sense of which parking strategies are successful and which are not. Here, the best solution is to assign *overlapping representations* to similar events — this overlap allows you to represent what all of your successful parking episodes have in common, and it allows the effects of these multiple episodes to accumulate on a common set of connection weights between neural units. Also, it is necessary to learn using a *small learning rate* — each experience should slightly adjust the weights such that, in the end, your parking strategy representation is roughly an average of all of your parking experiences.

It should be clear from this discussion that learning about specifics and learning about generalities are incompatible in neural networks — the two types of learning require different kinds of representations, and different learning rates. According to the CLS framework, we avoid having to make a tradeoff between these different functional demands by using two specialized learning

systems:

- The hippocampus, which is specialized for rapidly memorizing specific events.
- The neocortex, which is specialized for slowly learning about statistical regularities in the environment.

The hippocampus assigns distinct (*pattern separated*) representations to stimuli, thereby allowing it to learn rapidly without suffering catastrophic interference. In contrast, cortex assigns similar representations to similar stimuli; use of overlapping representations allows cortex to represent the shared structure of events, and therefore makes it possible for cortex to generalize to novel stimuli as a function of their similarity to previously encountered stimuli.

We have built hippocampal and neocortical networks that incorporate key aspects of the biology of these structures, and instantiate the complementary learning systems principles outlined above. In the following two sections, we describe the architecture of these networks, and how we have applied the networks to recognition memory. For a more detailed treatment of the neurobiological and functional constraints incorporated into the two networks, we refer the reader to our prior publications (O'Reilly & Rudy, 2001; McClelland et al., 1995;



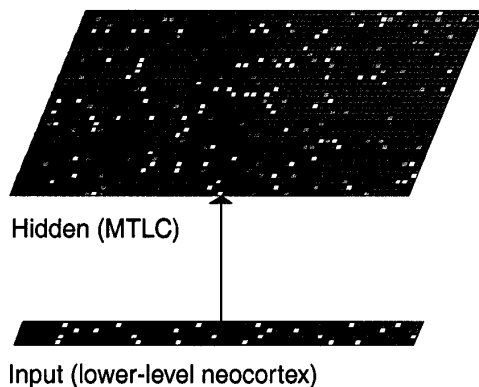


Figure 3: Diagram of the cortical network. The cortical network consists of two layers, an input layer (corresponding to “lower” cortical regions that feed into MTLC) and a hidden layer (corresponding to MTLC). Units in the hidden layer compete to encode (via Hebbian learning) regularities that are present in the input layer.

O’Reilly & McClelland, 1994; O’Reilly & Munakata, 2000; O’Reilly, Norman, & McClelland, 1998).

Both the hippocampal and neocortical networks were constructed using the *Leabra* model (O’Reilly & Munakata, 2000; O’Reilly, 1998, 1996), which brings together several widely-accepted characteristics of neural computation. These include Hebbian LTP/LTD (long-term potentiation/depression) and inhibitory competition between neurons, which are emphasized in the present model. Error-driven learning is part of the *Leabra* framework, but it was not incorporated in the simulations reported here. In the *Leabra* model, Hebbian LTP is implemented by strengthening the connection (weight) between two units when both the sending and receiving units are active together; Hebbian LTD is implemented by weakening the connection between two units when the receiving unit is active, but the sending unit is not (heterosynaptic LTD). Inhibitory competition is implemented in *Leabra* using a *k*-winners-take-all (kWTA) algorithm, which sets the amount of inhibition for a given layer such that *at most* *k* units in that layer have significant activation. The details of the *Leabra* algorithm are discussed at length in O’Reilly and Munakata (2000), and key aspects of the algorithm are summarized in Appendix A.

### The Cortical Model

The cortical network is composed of two layers, *input* (which represents the activation patterns of cortical areas that feed into the MTLC), and *hidden* (corresponding to MTLC) (Figure 3). The basic function of the

model is for the hidden layer to encode regularities that are present in the input layer; this is achieved through the Hebbian learning rule. To capture the idea that the input layer represents many different cortical areas, it consists of 24 10-unit *slots*, with 1 unit out of 10 active in each slot. Thus, each slot represents a different cortical area, roughly speaking. The hidden (MTLC) layer consists of 1920 units, with 10% activity (i.e., 192 of these units are active on average for a given input). The input layer is connected to the MTLC layer via a partial feedforward projection where each MTLC unit receives connections from 25% of the input units. When items are presented at study, these connections are modified via Hebbian learning.

Input patterns were constructed from prototypes by randomly selecting a new feature value (possibly identical to the old feature value) for a random subset of slots. The number of slots that were flipped (i.e., given a new value) when generating items from the prototype is a model parameter — increasing the number of slots that are flipped decreases the average overlap between items. When all 24 slots are flipped, the resulting item patterns have 10% overlap with one another (i.e., exactly as expected by chance in a layer with a 10% activation level). Thus, with input patterns one can make a distinction between *prototypical* features of those patterns, which have a relatively high likelihood of being shared across input patterns, and non-prototypical, *item-specific* features of those patterns (generated by randomly flipping slots) which are relatively less likely to be shared across input patterns. Prototype features can be thought of as representing both high-frequency item features (e.g., if you study pictures of people from Norway, most people have blond hair) as well as contextual features that are shared across multiple items in an experiment (e.g., the fact that all of the pictures were viewed on a particular monitor in a particular room on a particular day). Some simulations involve more complex stimulus construction, as described where applicable.

### Familiarity as Sharpening

To apply the cortical model to recognition, we exploit the fact that — as items are presented repeatedly — their representations in the MTLC layer become *sharper* (Figure 4). That is, novel stimuli weakly activate a large number of MTLC units, whereas familiar (previously presented) stimuli strongly activate a relatively small number of units. Sharpening occurs because Hebbian learning specifically tunes some MTLC units to represent the stimulus. When a stimulus is first presented, some MTLC units, by chance, will respond more strongly to the stimulus than other units; these units get tuned by Hebbian learning to respond even more strongly to the item then next time it is presented, and these strongly ac-

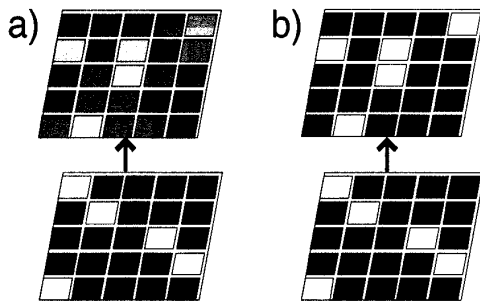


Figure 4: Illustration of the sharpening of hidden (MTLC) layer activation patterns in a miniature version of our cortical model. (a) shows the network prior to sharpening; MTLC activations (more active = lighter color) are relatively undifferentiated. (b) shows the network after Hebbian learning and inhibitory competition produce sharpening; a subset of the units are strongly active, while the remainder are inhibited. In this example, we would read out familiarity by measuring the average activity of the  $k = 5$  most active units.

tive units start to inhibit units that are less strongly active.

This sharpening dynamic in our model is consistent with neural data on the effects of repeated presentation of stimuli in cortex. Electrophysiological studies show that some neurons that initially respond to a stimulus exhibit a lasting decrease in firing, while other neurons that initially respond to the stimulus do not exhibit decreased firing (e.g., Brown & Xiang, 1998; Li, Miller, & Desimone, 1993; Xiang & Brown, 1998; Miller, Li, & Desimone, 1991; Rolls, Baylis, Hasselmo, & Nalwa, 1989; Riches, Wilson, & Brown, 1991). According to our model, this latter population consists of neurons that were selected (by Hebbian learning) to represent the stimulus, and the former population consists of neurons that are being forced out of the representation via inhibitory competition.

To index representational sharpness in our model — and through this, stimulus familiarity — we measure the average activity of the MTLC units that win the competition to represent the stimulus. That is, we take the average activation of the top  $k$  (192 or 10% of the MTLC) units computed according to the kWTA inhibitory competition function. This “activation of winners” (*act.win*) measure increases monotonically as a function of how many times a stimulus was presented at study. In contrast, the simpler alternative measure of using the average activity of *all* units in the layer is not guaranteed to increase as a function of stimulus repetition — as a stimulus becomes more familiar, the winning units become more active, but losing units become less active (due to inhibition from the winning units); the net effect

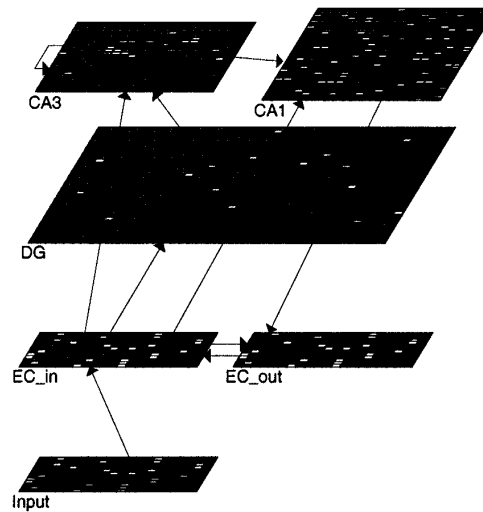


Figure 5: Diagram of the hippocampal network. The hippocampal network links input patterns in entorhinal cortex (EC) to relatively non-overlapping (pattern-separated) sets of units in region CA3; recurrent connections in CA3 bind together all of the units involved in representing a particular EC pattern; the CA3 representation is linked back to EC via region CA1. Learning in the CA3 recurrent connections, and in projections linking EC to CA3 and CA3 to CA1, makes it possible to recall entire stored EC patterns based on partial cues. The dentate gyrus (DG) serves to facilitate pattern separation in region CA3; see O'Reilly and McClelland (1994) for more details.

is therefore a function of the exact balance between these increases and decreases.

Although we will be using *act.win* in the simulations reported below, we do not want to make a strong claim that *act.win* is the way that familiarity is read out from MTLC. It is the most convenient and analytically tractable way to do this in our model, but there are other ways of reading out familiarity that might be employed by the brain (e.g., the time it takes for activation to spread through the network). We discuss this issue later in the general discussion section.

### The Hippocampal Model

We have developed a “standard model” of the hippocampus (O'Reilly et al., 1998; O'Reilly & Munakata, 2000; O'Reilly & Rudy, 2001; Rudy & O'Reilly, 2001) that implements widely-accepted ideas of hippocampal function (Hebb, 1949; Marr, 1971; McNaughton & Morris, 1987; Rolls, 1989; O'Reilly & McClelland, 1994; McClelland et al., 1995; Hasselmo, 1995).

In the brain, entorhinal cortex (EC) is the interface between hippocampus and neocortex; superficial layers of entorhinal cortex send input to the hippocampus, and deep layers of entorhinal cortex receive output from the hippocampus (see Figure 1). Correspondingly, our model subdivides EC into an EC.in layer that sends input to the hippocampus and an EC.out layer that receives output from the hippocampus. The basic function of the hippocampal model is to store patterns of EC.in activity, in a manner that supports subsequent recall of these patterns on EC.out.

In the hippocampal model, the input layer — which is configured identically to the cortical-model input layer — serves to impose a pattern of activation on EC.in via fixed, 1-to-1 connections; from there, activity spreads into the hippocampus. The three basic computational structures in the hippocampus are:

- The feedforward pathway from the entorhinal cortex input (EC.in) to area CA3 (via the dentate gyrus, DG), which produces pattern-separated representations of new memories in CA3 that are stored via Hebbian weight changes. These representations are *conjunctive*, in that they bind together disparate stimulus elements into a unitary representation. Both the pattern separation and conjunctivity effects arise from the use of *sparse representations* (where relatively few units are active for a given stimulus) in CA3 and especially DG (Marr, 1971; O'Reilly & McClelland, 1994).
- Recurrent connectivity within CA3, which binds together the units participating in a given representation. This is primarily important for recalling previously stored memories via *pattern completion*, whereby a partial input pattern reactivates the original CA3 representation.
- Area CA1, which translates between the CA3 encoding and the EC input/output representation. Thus, when pattern completion occurs in CA3, it subsequently generates activation patterns over the output (deep) layers of EC via CA1.

Figure 5 shows the structure of the model, and an example activation pattern. Table 1 shows that the model layers are roughly proportionately scaled based on the anatomy of the rat, but the activation levels are generally higher (less sparse) to obtain sufficient absolute numbers of active units for reasonable distributed representations given the small total number of units. These activity levels are enforced by setting appropriate  $k$  parameters in the Leabra kWTA inhibition function. Only Hebbian learning is used because it is sufficient for simple information storage.

Area	Rat		Model	
	Neurons	Activity (pct)	Units	Activity (pct)
EC	200,000	7.0	240	10.0
DG	1,000,000	0.5	1600	1.0
CA3	160,000	2.5	480	4.0
CA1	250,000	2.5	640	10.0

Table 1: Rough estimates of the size of various hippocampal areas and their expected activity levels in the rat, and corresponding values in the model. Rat data from Squire et al., 1989; Boss et al., 1987; Boss et al., 1985; Barnes et al., 1990.

In summary: The hippocampus supports recall via learning that occurs at study in connections linking EC to CA3 to CA1, and in the recurrent CA3 connections. When a previously studied EC.in pattern (or a subset thereof) is presented to the hippocampal model, the model is capable of reactivating the entire CA3 pattern corresponding to that item via strengthened weights in the EC-to-CA3 pathway, and strengthened CA3 recurrences. Activation then spreads from the item's CA3 representation to the item's CA1 representation via strengthened weights, and — from there — to the item's EC.out representation. In this manner, the hippocampus manages to retrieve a complete version of the studied EC pattern in response to a partial cue. In contrast, because of pattern separation, even partially novel stimuli tend to activate CA3 units that were not strongly linked to CA1 at study. As such, activity does not spread from CA3 to CA1, and recall does not occur. Even if the EC.in activity pattern corresponds to two components that were studied, but not together (see the *Associative Recognition* section, below), the conjunctive nature of the CA3 representations will minimize the extent to which recall occurs. Thus, there is a kind of "floor effect" on recall, whereby the weights linking an item's CA3 representation to CA1 have to be strengthened above a certain threshold before any recall occurs; studied inputs cross this threshold but nonstudied inputs rarely do.

#### The Hippocampal Recall Measure

To apply the hippocampal model to recognition, we exploit the fact that studied items tend to trigger recall (of the item itself), more so than lure items. Thus, a high level of match between the test probe (presented on the EC input layer) and recalled information (activated over the EC output layer) constitutes evidence that an item was studied. Also, we exploit the fact that lures sometimes trigger recall of information that *mismatches* the recall cue; for example, in the plurality-recognition experiment described earlier, switched-plurality lures sometimes trigger recall of the corresponding studied item (e.g., the test cue "RAT" may cause a subject to recall "I studied RATS, not RAT"). Thus, mismatch between recalled information and the test probe tends to

indicate that an item was not studied.

For each test item, we generate a *recall score* using the formula:

$$(match - mismatch) / (numslots) \quad (1)$$

where *match* is the number of recalled features (on EC\_out) that match the cue (on EC\_in), and *mismatch* is likewise the number that mismatch; *numslots* is a constant that reflects the total number of feature slots in EC (24, in these simulations).

One should appreciate that equation 1 is not the only way to apply the hippocampal model to recognition. For example, instead of looking at recall of the test cue itself, we could attach contextual tags to items at study, leave these tags out at test, and measure the extent to which items elicit recall of contextual tags. Also this equation does not incorporate the fact that recall of *item-specific* features (i.e., features unique to particular items in the item set) is more diagnostic of study status than recall of *prototypical* features — if all items in the experiment are fish, recall of prototypical fish features (e.g., “I studied fish”) in conjunction with a test item does not mean that you studied this particular item. Furthermore, the extent to which mismatch should be weighted in the decision rule will vary according to the structure of the experiment. For example, mismatching plurality recall is more diagnostic in an experiment where either the singular or plural form of a word is studied (but not both) than in an experiment where a given word might be studied in both singular and plural form. We selected the *match - mismatch* rule because it is a simple way to reduce the vector output of the hippocampal model to a scalar that correlates with the study status of test items. Assessing the optimality of this rule, relative to other rules, and exploring ways in which different rules might be implemented neurally, are topics for future research.

### Simulation Methodology

Our initial simulations involve a side-by-side comparison of the cortical and hippocampal networks receiving the exact same input patterns. This method allows us to analytically characterize differences in how these networks respond to stimuli. A shortcoming of this “side-by-side” approach is that we can not explore direct interactions between the two systems. To remedy this shortcoming, we will also present simulations using a *combined model* where the cortical and hippocampal networks are connected in serial (such that the cortical network that computes familiarity serves as the input to the hippocampal network) — this arrangement accurately reflects how cortex and hippocampus are arranged in the brain.

In our recognition simulations, the cortical and hippocampal models were (separately) given a list of items to learn, followed by a recognition test in which the models had to discriminate between 10 studied *target* items and 10 nonstudied *lure* items. No learning occurred at test. Unless otherwise specified, all of our recognition simulations used the same set of parameters (hereafter referred to as the *basic* parameters; these parameters are described in detail in Appendix B). In our basic-parameter simulations, we used a 20-item study list, and the average amount of overlap between items was 20% — 20% overlap was achieved by starting with a 24-slot prototype pattern, and then generating items by randomly selecting new feature values for 16 randomly selected slots. To facilitate comparison between the models, we used hippocampal and cortical parameters that yielded roughly matched performance across the two models for both single-probe (yes-no, or *YN*) and forced-choice (*FC*) recognition tests. We matched performance in this way to alleviate concerns that differential effects of manipulations on hippocampal recall and MTLC familiarity are attributable simply to different overall levels of performance in the two networks. However, this matching does not constitute a strong claim that hippocampal and cortical performance are — in reality — matched when overlap equals 20% and study list length equals 20.

To simulate yes-no (*YN*; single-probe) recognition performance, we computed hits and false alarms in the two models by applying thresholds to the MTLC familiarity and hippocampal recall measures, respectively. For the cortical model, we set an unbiased threshold for each simulated subject by computing the average familiarity scores associated with studied and lure items (respectively) and then placing the familiarity threshold exactly between the studied and lure means. For the hippocampal model, we took a different approach to threshold setting; as discussed in the *ROC Curves* section below, it is possible to set a *high recall threshold* that is sometimes crossed by studied items, but never crossed by lures. We assume that subjects are aware of this fact (i.e., that high amounts of recall are especially diagnostic of having studied an item) and set a recall threshold that is high enough to avoid false recognition. Accordingly, in our basic-parameter simulations, we used a fixed, relatively high threshold (recall = .40); this value was chosen because — assuming other parameters are set to their “basic” values — it is sometimes exceeded by studied items, but never by lures (unless lures are constructed to be similar to specific studied items; see *Simulation 4* for more details).

We used *d'* (computed on individual subjects' hit and

false alarm rates) to index YN recognition sensitivity.<sup>1</sup> In most of the simulations reported below, the qualitative pattern of  $d'$  results is not contingent on threshold placement (i.e., the results would be qualitatively the same if we used an unbiased as opposed to a high threshold for recall). In the few situations where recall threshold placement does matter, we will note this fact and explain how threshold placement impacts performance.

The goal of the simulation work presented here is to establish key qualitative properties of the two networks, using (whenever possible) a fixed set of underlying parameters. The model has not yet reached the point where it can provide quantitative fits to recognition data, because — as discussed in the *Variability and Scaling Effects* section of the paper — the number of units in our models is very small relative to the actual number of neurons in MTLC and the hippocampus, and there are sources of variance present in our (relatively) small-scale models we run that would not be present in a brain-sized model. Furthermore, there are other sources of variance (e.g., variability in how well different items are encoded) that are not yet implemented in the model. Thus, our models' predictions regarding how manipulations affect variance will not necessarily coincide with the predictions of larger models that incorporate other sources of variance like encoding variability, and variability due to pre-experimental exposure to list items.

All of the simulation results reported in the text of the paper are significant at  $p < .001$ . In graphs of simulation results (starting with Figure 10 below), error bars indicate the standard error of the mean, computed across simulated subjects. When error bars are not visible, this is because they are too small relative to the size of the symbols on the graph (and thus are covered by the symbols).

Finally, a point of rhetorical clarification. We use the terms “recall” and “familiarity” to describe the respective contributions of the hippocampus and MTLC to recognition memory, because these terms are heuristically useful. The hippocampal contribution to recognition is “recall” insofar as it involves retrieval of specific studied details. We use “familiarity” to describe the MTLC contribution insofar as — like the other mathematical models discussed earlier — the MTLC signal is a scalar that tracks the global match or similarity of the test probe to studied items. However, the reader should keep in mind that the contributions of MTLC and the hippocampus to recognition (as set forth by the Complementary Learning

<sup>1</sup>To avoid problems with  $d'$  being undefined when hit or false alarm rates equal 0 or 1, we adjusted hit and false alarm rates using the correction suggested by Snodgrass and Corwin (1988) prior to computing  $d'$ :  $P = (n + .5)/(N + 1)$ , where  $n$  = the number of “old” responses,  $N$  = the total number of items, and  $P$  = the corrected percent “old” value.

Pattern Separation in MTLC and the Hippocampus

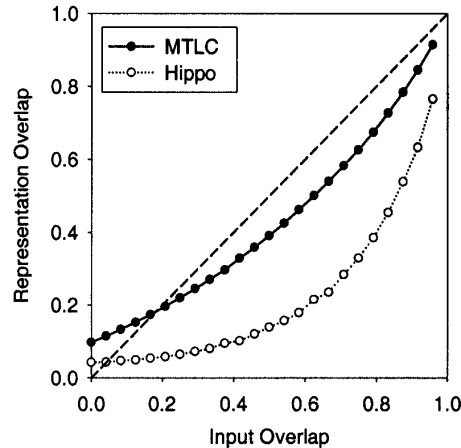


Figure 6: Results of simulations exploring pattern separation in the hippocampal and cortical models. In these simulations, we created pairs of items and manipulated the amount of overlap between paired items. The graph plots the amount of input-layer overlap for paired items versus: 1) CA3 overlap in the hippocampal model, and 2) MTLC overlap in the cortical model. In this graph, all points below the diagonal (dashed line) indicate pattern separation (i.e., representational overlap < input overlap). The hippocampal model showed a strong tendency towards pattern separation (CA3 overlap < input overlap); the cortical model showed a smaller tendency towards pattern separation (MTLC overlap was slightly less than input overlap).

Systems model) do not map perfectly onto existing ideas regarding how recall and familiarity contribute to recognition. Indeed, we will systematically delineate how the CLS model coincides with, and deviates from, existing dual-process frameworks.

## Part I: Basic Network Properties

Simulations in this section address basic properties of the cortical and hippocampal networks, including differences in their ability to assign distinct (“pattern-separated”) representations to input patterns, and differences in their operating characteristics. We also discuss sources of variability in the two networks.

### Simulation 1: Pattern Separation and Blending

We begin the process of establishing the critical differences between the MTLC familiarity signal and the

hippocampal recall signal by focusing on pattern separation — the ability to assign relatively non-overlapping representations to input patterns. As we will see, many of the differences between hippocampally- and cortically-driven recognition discussed later in the paper can be traced back to differences in the two networks' pattern separation ability. We ran simulations exploring pattern separation in the two networks; in these simulations, we created pairs of items and manipulated the amount of overlap between paired items. We had the network "study" one item from each pair, and the other item was presented at test. Figure 6 plots the amount of *input-layer* overlap between paired study and test items, against both hippocampal and cortical overlap for paired items. Hippocampal overlap was measured as % overlap between paired items in CA3; cortical overlap was measured in the hidden (MTLC) layer. Pattern separation is evident when representational (CA3 or MTLC) overlap is less than the input overlap. As expected, there is much greater pattern separation in the hippocampal model compared to the cortical model; this is due primarily to the greater levels of sparseness in the hippocampal representations (see O'Reilly & McClelland, 1994 for a thorough analysis). The cortical model exhibits a relatively mild level of pattern separation; MTLC overlap generally tracks the overlap between the input patterns. These simulations also show that the hippocampus' ability to assign distinct representations to stimuli is *limited* — as overlap between input patterns increases, hippocampal overlap eventually increases above floor levels (although it always lags behind input pattern overlap).

An important consequence of hippocampal pattern separation is that recall of *blends* of multiple studied items is rare. Generally speaking, features will be recalled together only if they were studied together. The only exception to this rule occurs when the average amount of overlap between input patterns is high. In this situation, pattern separation starts to break down, and frequently-occurring prototype features sometimes intrude into the recall vector (supplanting item-specific features).

We ran a series of simulations using the hippocampal model to explore the model's robustness to blending. In these simulations, we had the model study 20 input patterns and re-presented these studied patterns at test; we also manipulated the average amount of overlap between input patterns (from 20% to 50%). In simulations with 20% input overlap, the rate of blending was extremely low: on more than 98% of trials where the test cue triggered recall, the recall vector was a partial version of a specific studied item (as opposed to a blend of multiple items). However, with higher levels of overlap, the rate of blending increased. This increase is almost entirely attributable to prototype features intruding into the re-

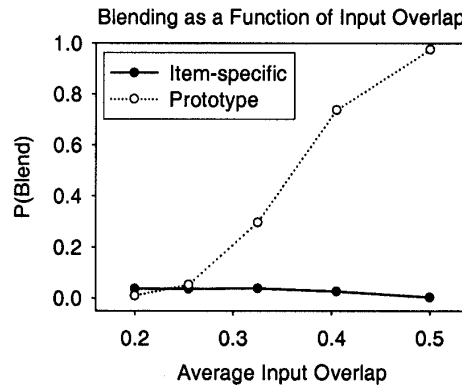


Figure 7: Probability of *prototype* blending (i.e., intruding a prototype feature into recall of some other pattern) and *item-specific* blending (i.e., intruding a non-prototypical feature into recall of some other pattern), as a function of the average level of overlap between input patterns. Prototype blending increases as overlap increases, but item-specific blending stays close to floor.

call vector (Figure 7). The probability of *item-specific* blending (i.e., recall of item-specific features that did not occur together at study) remained very low, even with high levels of overlap. The small amount of item-specific blending that did occur is probably attributable to imperfect (pre-experimental) learning of the mapping between CA1 and EC, rather than blending per se.

### Simulation 2: ROC Curves

Another way to characterize how cortical and hippocampal contributions to recognition differ is to graph the operating characteristics of these signals using ROC curves (Green & Swets, 1966). ROC curves plot hits vs. false alarms while varying the recognition threshold; the area under the ROC increases as a function of recognition sensitivity. These curves have added importance because the *dual-process signal-detection* framework of Yonelinas & Jacoby (Yonelinas & Jacoby, 1996; Jacoby et al., 1997) makes specific assumptions about the operating characteristics of recall and familiarity in order to measure the respective contributions of these processes to behavioral recognition performance. Specifically, these assumptions are:

- Familiarity is an equal-variance signal-detection process: studied (signal) and lure (noise) familiarity are both normally distributed; the two distributions have equal variance and overlap extensively.
- Recall is a high-threshold process: recall is all-

or-none; studied items are sometimes called "old" based on recall, but lure items are never called "old" based on recall.

The assumption that studied and lure familiarity are normally distributed implies that the ROC curve for familiarity (generated by sweeping a threshold across these distributions) should be curvilinear. The normality assumption also implies that it will be impossible to completely eliminate familiarity false alarms without also eliminating familiarity hits; as such, the Y-intercept of the ROC (i.e., the hit rate when false alarms = 0) should be zero. Finally, the assumption that the studied and lure familiarity distributions have equal variance implies that the ROC should be symmetrical.

The assumption that recall is a high-threshold process implies that the ROC curve for recall should have a positive Y-intercept equal to the probability of recalling a studied item (i.e., when subjects do not guess, the hit rate equals the probability of recalling a studied item, and the false alarm rate equals zero). High-threshold theory posits that any recall false alarms are due to guessing on test trials where "true recall" does not occur; varying the probability of guessing yields a linear ROC curve (see Macmillan & Creelman, 1991 for more background on ROC curves).

To generate ROC curves, we swept a threshold across the *act.win* MTLC familiarity measure for the cortical network, and across the *match - mismatch* recall measure for the hippocampal network, recording the associated proportions of hits (for studied items) and false alarms (for lure items). Four levels of average input overlap were used, ranging from .2 to .5, and lures were *unrelated* to studied items (i.e., lures and studied items were randomly sampled from the same pool of patterns, such that the average overlap between studied items and lures was equal to the average overlap between different studied items). Results are plotted in Figure 8.

First, we consider the cortical ROC curves (Figure 8a). Regardless of overlap, the curves are smoothly curvilinear with a zero Y-intercept, consistent with the idea that MTLC familiarity is a classic signal detection process. Also, the curves are approximately symmetrical, which is consistent with Yonelinas' assumption that familiarity is an *equal-variance* signal-detection process. However, we should not infer too much from this finding of symmetry (and, by inference, equal variance), because — as discussed in the *Variability and Scaling Effects* section — there are sources of variance in this model that are not present in brain-sized models, and there are also sources of variance that are missing from this model.

Next, we consider the hippocampal ROC curves (Figure 8b). In the 20% (.2) input overlap condition, the hippocampal ROC has a high Y-intercept (around .8), and

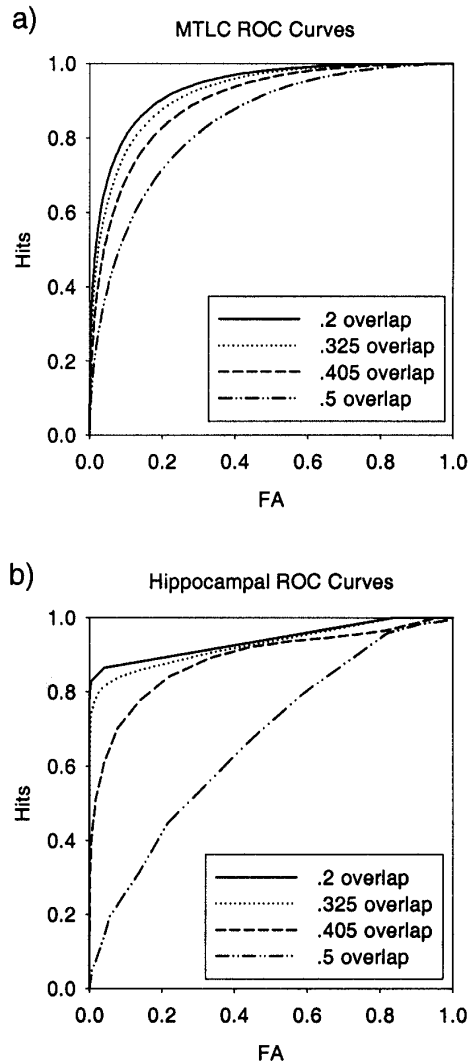


Figure 8: (a) ROC curves generated by sweeping a threshold across the *act.win* MTLC familiarity measure; (b) ROC curves generated by sweeping a threshold across the *match - mismatch* hippocampal recall measure. In both cases, input overlap was manipulated from .2 to .5. The MTLC ROC curves are curvilinear with Y-intercept = 0 regardless of input overlap; the hippocampal ROC is (mostly) linear with a positive Y-intercept for low overlap values; for higher overlap values, the hippocampal ROC is curvilinear, and the Y-intercept is closer to zero.

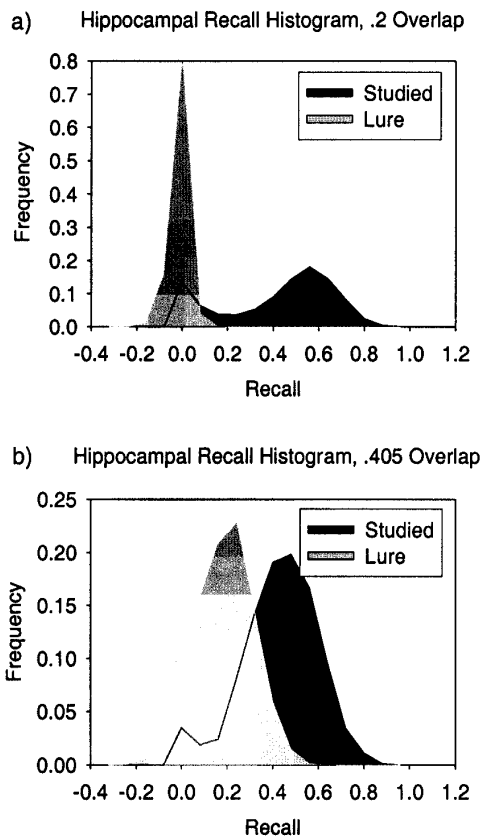


Figure 9: (a) histogram of the studied and lure hippocampal recall distributions for low input overlap (.2); (b) histogram of the studied and lure recall distributions for a higher input overlap value (.405).

the curve is generally linear but there is a small bend close to the Y axis. Similar behavior occurs for the .325 overlap case as well. Thus, for low-to-moderate levels of input overlap, hippocampal ROC curves are *approximately* consistent with the Yonelinas' assumption that recall is a high-threshold process. With higher levels of overlap, the hippocampal ROC is curvilinear, with a Y-intercept that is closer to zero. In other words, the hippocampus transitions from this approximate high-threshold behavior to signal-detection behavior (much like the cortical model) as a function of input overlap.

To gain further insight into the two different modes of hippocampal function, Figure 9 plots the underlying distributions of recall scores for studied items and lures (from which the ROC curves were constructed) for low (.2) and high (.405) average input overlap.

First, we will consider the low overlap condition (Figure 9a). The approximate linearity of the ROC in this condition occurs because the bulk of the lure recall distribution is located at the zero recall point; sweeping the threshold across the zero point causes a large spike in false alarms, and the ROC consequently jumps from a point very close to the Y axis (zero false alarms) to a point further away from the Y axis. The absence of strong lure recall in the 20% overlap condition is a consequence of hippocampal pattern separation and the thresholded nature of hippocampal recall, as discussed earlier when the hippocampal model was introduced. The ROC is not completely linear because the lure recall distribution is not completely located at the zero point; some lures trigger above-zero (but still low) levels of recall. The high Y-intercept of the ROC derives from the fact that the maximum recall score triggered by lures is much lower than the maximum recall score triggered by studied items; thus, it is possible to achieve a high hit rate without making any false alarms. The maximum amount of matching recall triggered by lures is limited by two factors. First, the hippocampus will not recall *nonstudied* features of lures because of the thresholded nature of recall — generally speaking, the hippocampus will only recall a feature if weights to that feature's CA1 representation (from CA3) were strengthened at study. The second limiting factor is the property, mentioned earlier, that blending is rare: Item-specific features of lures that did not occur together at study will not be recalled together at test.

The studied recall distribution is bimodal because of nonlinear attractor dynamics in region CA3. If the recurrent weights linking active units in CA3 are sufficiently strong, this generates positive feedback effects that amplify CA3 activity, thereby boosting recall. Most studied items benefit from these positive feedback effects but, because of variability in initial weight values, some studied items do not have weights strong enough to yield positive feedback effects. These items only weakly activate CA3 and are poorly recalled, thereby accounting for the extra "peak" at recall = 0.

Next, we consider the high overlap condition (Figure 9b). In this condition, both studied items and lures trigger large amounts of recall, such that the studied and lure recall distributions are roughly normal and overlap extensively. High levels of lure recall occur in the high-overlap condition because of *pattern separation failure* in the hippocampus; as documented in Simulation 1 (Figure 6), the hippocampus loses its ability to assign distinct representations to input patterns when overlap between inputs is very high. In this situation, the same CA3 units — the units that are most sensitive to frequently-occurring prototype features — are activated again and again by studied patterns, and these units acquire very



strong weights to the representations of prototype features in CA1. When items are presented at test, they activate these "core" CA3 units to some extent (regardless of whether or not the test item was studied), and activation spreads very quickly to CA1, leading to possibly erroneous recall of prototype features. This phenomenon, whereby increasing overlap increases erroneous recall of prototype features, was documented in our exploration of feature blending (Figure 7). When pattern separation failure occurs, prototype recall (which is relatively non-discriminative, because it is triggered by both studied items and lures) swamps recall of more discriminative, item-specific features, thereby boosting overlap between the studied and lure distributions and lowering overall discriminability.

In summary: The ROC results show that MTLC familiarity is a standard signal detection process, as assumed by Yonelinas' dual-process model. Hippocampal recall has two modes of operation: When input patterns have low-to-moderate average overlap, the hippocampus exhibits *approximate high-threshold* behavior — studied items trigger recall (of item-specific and prototype features) but lures trigger virtually no recall; in this condition, our model's ROC curves resemble the high-threshold ROC curves predicted by Yonelinas in some ways (a high Y-intercept) but not in others (the curves are not strictly linear). In contrast, when input patterns have high average overlap, recall functions as a standard signal-detection process — both studied items and lures trigger varying degrees of prototype recall, leading to curvilinear ROCs with a zero Y-intercept.

### Simulation 3: Variability and Scaling Effects

Recognition performance involves detecting the presence of a variable memory signal against a background of noise. Many distinct forms of variability can affect recognition performance; we need to carefully delineate which of these sources of variability are present in our models, because — as we will show later — different forms of variability have different implications for recognition performance. We show here that the primary source of variability in our models is *sampling variability*: variation in how well, on average, neural units are connected to (sampled by) other neural units in the network.<sup>2</sup>

However, we also show that the magnitude of sam-

pling variability is an inverse function of the size of the networks, such that in a network scaled to the approximate size of the human brain, this form of variability would likely be negligible. Therefore, we conclude that other forms of variability must be at play in the human brain; we show how these other forms of variability can be captured in our models and discuss the implications of including other forms of variability.

### Sampling Variability

Sampling variability arises in the cortical network for several reasons. First, each unit in MTLC only has connections to a randomly selected subset of input units (25%); this partial connectivity helps MTLC form specialized representations of the input space (i.e., they come to represent some input features but not others). It also has the consequence that, by chance, some input features will be sampled better (i.e., they will have more connections to MTLC) than other input features. Furthermore, connections start with random weight values (sampled from a uniform distribution with a range of .25 around a mean of .5), and this adds to the variability in how well different input features are represented in MTLC. Thus, input patterns that by chance happen to be well-sampled tend to trigger higher familiarity scores than input patterns that are poorly sampled.

Sampling variability also comes into play in several places in the hippocampal network. For example, there is variability in the (pre-learning) strength of weights linking CA3 to CA1. If the weights between the CA3 representation of an item and its CA1 representation happen to be small prior to learning, this will hinder subsequent recall (i.e., if they weights start small, they may be too small even after learning to support recall). Also, there is variability in the strength of the recurrent connections between CA3 units. If the CA3 units activated by an item happen, by chance, to be densely interconnected, this will increase positive feedback between CA3 units activated by the item at test, leading to increased overall activation of CA3 units and better recall (and vice-versa for less densely interconnected units).

We can perform some basic calculations to show that sampling variability decreases with increasing network size. Sampling variability in the cortical model is primarily attributable to variability in how many times (across all MTLC units) an input unit is sampled. The probability that an MTLC unit will sample a given input unit (assuming 25% connectivity) is .25, so the total number of times that an input unit is sampled, across all MTLC units, is binomially distributed with parameters  $p = .25$ ,  $q = 1 - p = .75$ ,  $n =$  number of MTLC units. Therefore, the variance of the *proportion* of MTLC units sampling

<sup>2</sup>Note that our use of the term "sampling variability" differs from how other modelers have used this term. In our model, sampling variability is a function of variability in the initial values assigned to weights in the network. Other models use sampling variability to refer to variability in which item features are presented to the model at study and test (Atkinson & Estes, 1963), or variability in which memory trace is retrieved at test (Gillund & Shiffrin, 1984).

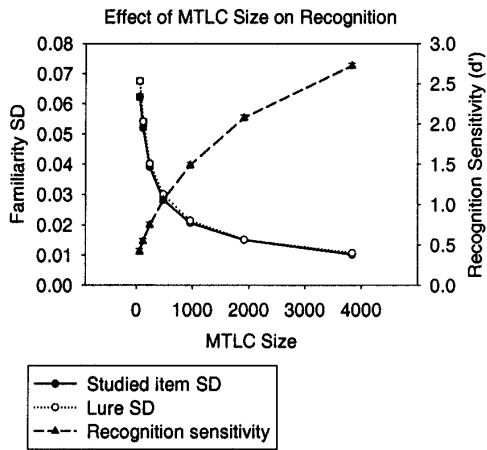


Figure 10: Graph plotting how the size of the MTLC layer affects the standard deviation (SD) of the familiarity signal triggered by studied items and lures, as well as overall recognition performance. As MTLC size increases, the SD of the studied and lure familiarity distributions decreases, and overall recognition performance increases.

a given input unit is:

$$\text{var}(S) = \frac{pq}{n} \quad (2)$$

From this equation, it is evident that variability in how often (proportionally) an input unit is sampled will decrease as the number of hidden units  $n$  increases. In a brain-sized network, with millions of units (instead of the hundreds in our model), sampling variability will be far too small to affect recognition performance.

We ran some simulations to illustrate this effect, in which we manipulated MTLC (hidden layer) size and recorded the standard deviation of the studied and lure familiarity distributions; we also recorded overall recognition sensitivity ( $d'$ ). Figure 10 shows the results; increasing MTLC size lowers the SD of the studied and lure familiarity distributions; consequently, there is less overlap between the studied and lure distributions, and recognition sensitivity increases.

Increasing network size also reduces sampling variability in the hippocampus — here, the key is that increasing network size increases the number of weights that are involved in storing a particular pattern; increasing the number of weights makes the *average* of these weights less variable, thereby decreasing the odds that — by chance — these weights will be too weak on average to support recall. Figure 11 shows the results of simulations exploring the effect of CA3 size on recall of studied items in the hippocampal model; as CA3 size in-

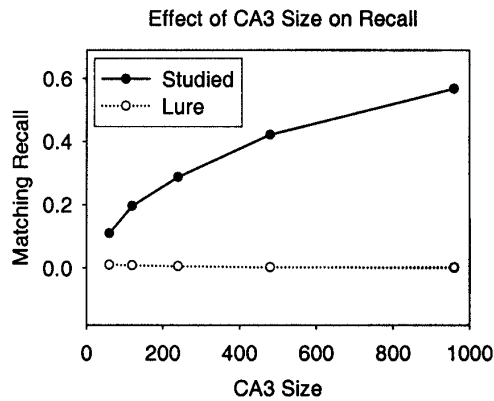


Figure 11: Graph plotting how the size of CA3 affects recall performance in the hippocampal network. As CA3 size increases, the amount of matching recall triggered by studied items increases. The amount of matching recall triggered by lures stays at floor.

creases, the mean amount of matching recall triggered by studied items increases. Asymptotically, as CA3 gets large enough, studied recall approaches ceiling, and consequently variability in studied recall approaches floor.

In short, stimulus sampling variability, though dominant in our small-scale simulations, should not be a major factor in brain-sized networks. Therefore, we need to look for other sources of variability in understanding human performance.

### Other Sources of Variability

A potentially important source of variability in recall and familiarity scores is variability in how well stimuli are encoded at study. This kind of *encoding variability* can arise, for example, if subjects' attention fluctuates over the course of an experiment — some items will be encoded more strongly than others, leading to higher recall and familiarity scores at test.

An important property of encoding variability, which is not true of sampling variability, is that it affects studied items and similar nonstudied items (*related lures*) in tandem; that is, encoding fluctuations that boost the memory signal triggered by a studied item will also boost the memory signal triggered by lures that are similar to that studied item (e.g., if "cat" is encoded so as to be especially familiar, the related lure "cats" will also be highly familiar). In contrast, sampling variability operates independently on each input feature. Variability in the sampling of *shared* features of studied items and related lures will always push the memory signal triggered by studied items and similar lures in the same direction,

but variability in the sampling of *discriminative* (non-shared) features of studied items and related lures can push studied and lure memory in different directions. In small networks where sampling variability is the dominant source of variance, unshared noise associated with sampling of discriminative features of overlapping stimuli counteracts much of the shared variability in memory scores triggered by these items. We will revisit this issue later, when we present simulations using related lures.

Another source of variability in recall and familiarity scores is variability in *pre-experimental exposure* to stimuli: Some stimuli have been encountered extensively prior to the experiment, in many different contexts; other stimuli are relatively novel; for evidence that pre-experimental presentation frequency affects recognition memory, see Dennis and Humphreys (2001). Variability in pre-experimental exposure (like encoding variability, but unlike sampling variability), reliably affects studied items and related lures in tandem.

Finally, in addition to variability in how much test items overlap with pre-experimental memory traces, there is also variability in how much items overlap with other items presented *in the experiment*; this kind of variability also affects studied items and related lures in tandem. Overlap-related variability is already present in the model, but its effect on performance is typically dwarfed by sampling variability. Consequently, variability in overlap should play a much larger role, proportionally, in larger networks with minimal sampling variability.

### *How Variability is Implemented in the Models*

Given that sampling variability is not likely to be a factor in human recognition memory performance, but it is dominant in our small-scale models, one might conclude that we should eliminate this source of variability and incorporate the more plausible sources just discussed. Unfortunately, this is not practical at present — models that are large enough to eliminate sampling variability cannot be feasibly run on available computational hardware. Furthermore, adding more variability on top of sampling variability in our small networks leads to poor performance, unless other steps are taken to compensate for increased variability (e.g., increasing the learning rate).

Nevertheless, it is relatively straightforward to incorporate other sources of variability, and we need to do so to make some important points later. For example, encoding variability can be implemented by randomly scaling the learning rate by a multiplicative factor on each study trial (simulating variations in attention), or we can randomly delete features from patterns when they are presented at study. To implement pre-experimental

variability in our model, we could pre-train experimental stimuli (and vary the amount of pretraining that each item receives) prior to the start of the actual “experiment”.

In summary, despite the limitations of our small networks, we can still use them to understand many important phenomena that do not depend on the exact source of variability; for those phenomena that do require specific forms of variability, we can simulate these phenomena by adding the appropriate form of variability on an “as needed” basis. However, because of the limitations discussed in this section, we will refrain from making strong predictions about variance in this paper.

## Part II: Applications To Behavioral Phenomena

The simulations in this part of the paper build on the basic results described earlier, by applying the models to a wide range of empirical recognition memory phenomena (e.g., how does interference affect recognition performance in the two models). Whenever possible, we will present data that speak to the model’s predictions.

### Simulation 4: Lure Relatedness and Test Format Interactions

Here, we explore the implications of the pattern separation differences between cortex and hippocampus (described in Simulation 1) and the implications of incorporating different kinds of variability in the models, using recognition simulations where we manipulate *lure relatedness* (i.e., the extent to which lure items resemble studied items).

As we saw in Figure 6, the hippocampus was better able to separate out the representations of overlapping inputs than the cortex; this finding implies that hippocampus should outperform cortex on tests where subjects have to discriminate between studied items and similar (overlapping) lures. However, our simulations — described below — also show that we can reduce this difference in performance by using a forced-choice (FC) test format instead of the yes-no (YN) format typically used. The FC format allows shared variability between studied items and closely related lures to be subtracted away, thereby boosting cortex’s ability to discriminate between studied items and similar lures.

### *YN Performance*

First, we present the results of YN recognition simulations where we varied target-lure similarity. We used our basic parameters, with 20% average overlap between

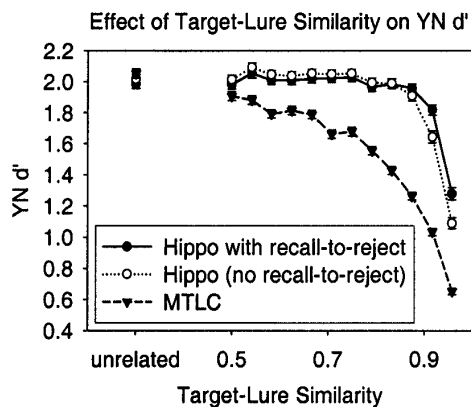


Figure 12: YN recognition sensitivity ( $d'$ ) in the two models, as a function of target-lure similarity. Target-lure similarity was operationalized as the proportion of input features shared by targets and corresponding lures; note that the average level of overlap between studied (target) items was held constant at 20%. These simulations show that the hippocampal model is more robust to increasing target-lure similarity than the cortical model; use of a recall-to-reject rule (where an item is called new if it triggers *any* mismatching recall) further benefits hippocampal performance when target-lure similarity is high.

studied items. For each studied (target) item, we created a corresponding lure item by taking the studied item and flipping a pre-specified number of slots; to vary target-lure similarity, we varied the number of slots that we flipped to generate lures (less flipping results in more overlap). For comparison, we also ran simulations with “unrelated” lures that were sampled from the same item pool as studied items. We computed recognition  $d'$  values for the two models in the standard manner (i.e., based on the *match* – *mismatch* hippocampal recall score, and the *act.win* MTLC familiarity score). In addition, we also explored the use of a “recall-to-reject” rule for the hippocampus that places a much stronger weight on mismatching recall. Whereas the *match* – *mismatch* rule weights matching and mismatching recall equally, the recall-to-reject rule posits that an item should be rejected if it produces *any* mismatching recall (otherwise, the recognition decision should be made based on whether the item triggers an above-threshold amount of matching recall). The recall-to-reject rule exploits the fact that mismatching recall is highly diagnostic of an item being new (nonstudied) in this simulation; related lures sometimes trigger mismatching recall but studied items virtually never do this. Figure 12 plots cortical  $d'$  scores and hippocampal  $d'$  scores (computed using both the *match* – *mismatch* rule and the recall-to-reject rule) as a function of target-lure similarity. Recognition per-

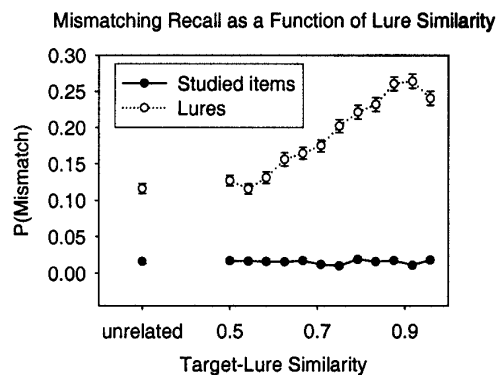


Figure 13: Plot of the probability that lures and studied items will trigger mismatching recall, as a function of target-lure similarity. This probability is close to floor for studied items; the probability that lures will trigger mismatching recall increases with increasing target-lure similarity.

formance based on MTLC familiarity gets steadily worse as lures become increasingly similar to studied items; by contrast, the hippocampus is relatively robust to the lure similarity manipulation —  $d'$  does not decrease appreciably until target-lure similarity approaches 90%. Figure 12 also shows that hippocampal performance is better for related lures when the “recall-to-reject” rule is used.

The cortical model results can be explained in terms of the fact that the cortical model assigns similar representations to similar stimuli — because the representations of similar lures (vs. dissimilar lures) *overlap more* with the representations of studied items, similar lures *benefit more* from learning that occurs at study. Thus, lure familiarity smoothly tracks target-lure similarity; increasing similarity monotonically lowers the target-lure familiarity difference, leading to decreased discriminability. In contrast, the hippocampal recall signal triggered by lures is stuck at floor until target-lure similarity > 60%, and lures do not start to trigger above-threshold (i.e., > .40) recall until target-lure similarity > 80%. This occurs because of hippocampal pattern separation and the thresholded nature of hippocampal recall — lures have to be very similar to studied items before they access enough strengthened weights to trigger recall.

The hippocampus also benefits from the fact that it can reject items based on mismatching recall. Figure 13 plots the probability that studied items and lures will trigger mismatching recall, as a function of target-lure similarity. This figure confirms the point, made earlier, that studied items virtually never trigger mismatching recall, but lures sometimes do; also, it shows that mismatching recall triggered by lures increases substan-

tially with increasing target-lure similarity. This increase in mismatching recall helps offset, to some degree, increased *matching* recall triggered by related lures. Hippocampal recognition performance is slightly better for related lures with the recall-to-reject rule vs. the standard rule because recall-to-reject assigns a higher weight to this (highly diagnostic) mismatch factor. With recall-to-reject, the only way that lures can trigger an old response in this case is if they trigger a large amount of *matching* recall but no *mismatching* recall. The odds of this happening are very low.

Importantly, the hippocampal model's robustness to the target-lure similarity manipulation (i.e., the fact that  $d'$  does not decrease until target-lure similarity approaches 90%) depends in part on our use of a high recall threshold. Because of the high threshold, false recognition is at floor from 60%-80% similarity even though the mean lure recall score is greater than zero in these conditions. Lowering the recall threshold from .4 to 0 would have the effect of moving up the point at which false recognition starts to occur (and  $d'$  starts to decrease) from around 90% similarity to around 60% similarity.

In summary: The model predicts that both hippocampal recall and MTLC familiarity should be able to support good performance on YN recognition tests with lures that are *unrelated* to studied items, but only the hippocampal recall signal can support good performance on YN recognition tests with *related* lures (i.e., lures with considerable feature overlap with studied items). This prediction is consistent with the view, expressed in several empirical papers, that recall is especially important for discriminating between studied items and very similar distractors (e.g., Hintzman et al., 1992; Rotello, Macmillan, & Van Tassel, 2000).

### FC Performance

We now show that the use of a forced-choice (FC) test format can improve the cortical network's recognition performance with related lures. In an FC test, subjects are simultaneously presented with a studied item and a lure, and are asked to select the studied item. The specific version of this test that boosts cortical performance involves pairing studied items with *corresponding* related lures (i.e., lures related to the paired studied item; for example, study "RAT", test "RAT" vs. "RATS").

The central insight as to why this format improves cortical performance with related lures is that, even though related lures trigger strong feelings of familiarity (because they overlap with the studied items), corresponding studied items will reliably be more familiar. Because performance in an FC test is based on the *difference* in familiarity between paired items, even small differences can drive good performance, as long as they

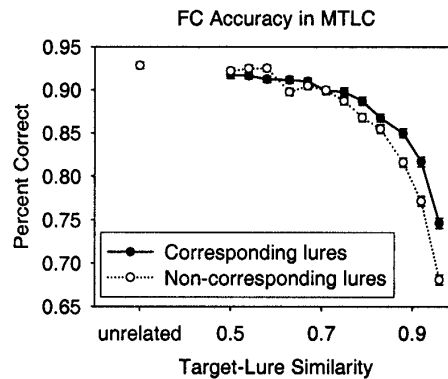


Figure 14: FC accuracy in the MTLC model as a function of target-lure similarity, using corresponding and non-corresponding FC testing. For high levels of target-lure similarity, FC performance is slightly better with corresponding lures vs. with non-corresponding lures.

are reliable.

Importantly, the reliability of the familiarity difference depends on where variability comes from in the model. As discussed earlier, some kinds of variability (e.g., differences in encoding strength and pre-experimental exposure) necessarily affect studied and related lure familiarity in tandem, whereas other kinds of variability (e.g., sampling variability) do not. When the former kind of variability predominates, the familiarity values of studied items and corresponding lures are highly correlated, and therefore their difference is reliable. When the sampling variability predominates, the studied-lure familiarity difference is less reliable.

More formally, the beneficial effect of using an FC test depends on *covariance* in the familiarity scores triggered by studied items and corresponding related lures (Hintzman, 1988, 2001). Specifically, FC recognition performance is a function of the variability in the difference in familiarity between studied items and paired lures:

$$Var(S - L) = Var(S) + Var(L) - 2 * Cov(S, L) \quad (3)$$

where  $S$  represents familiarity of studied items and  $L$  that of lures,  $Var$  is variance and  $Cov$  is covariance between  $S$  and  $L$ . Thus, the variance of the  $S - L$  familiarity difference is a function of the covariance — if covariance is high, the difference in familiarity between studied items and related lures will be highly reliable and FC performance will be good.

### The Cortical Model

To determine if the basic cortical model exhibits the covariance necessary to have a FC advantage for related

lures, we ran simulations using a paradigm introduced by Hintzman (1988). We compared FC performance with *corresponding* related lures (i.e., study A, B; test A vs. A', B vs. B', where A' and B' are lures related to A and B, respectively) to FC performance with *non-corresponding* lures (e.g., study A, B; test A vs. B'; B vs. A'). To the extent that there is covariance between studied items and corresponding lures, this will benefit performance in the corresponding lure condition relative to the non-corresponding lures. As shown in Figure 14, FC performance is higher with corresponding related lures than with non-corresponding lures — this replicates the empirical results obtained by Hintzman (1988) and shows that there is some covariance present in the basic cortical model.

To quantify the level of covariance underlying these results, we can compute the following ratio:

$$R = \frac{2 * Cov(S, L)}{Var(S) + Var(L)} \quad (4)$$

When  $R = 1$ , covariance will completely offset studied and lure variance, and the studied-lure familiarity difference will be completely reliable (i.e., variance = 0);  $R = 0$  means that there is no covariance. Averaging across the three highest target-lure similarity values (.96, .92, .88), the covariance ratio  $R = .31$  in the corresponding condition and  $R = -.03$  in the non-corresponding condition. Thus, the model exhibits roughly one third the maximal level of covariance possible.

Although the basic model can be said to qualitatively exhibit the FC advantage with corresponding related lures, this advantage is not quantitatively very large. This is because the dominant source of variability in the basic cortical model is sampling variability, which — as discussed above — does not reliably affect studied items and corresponding lures in tandem.

#### *Simulations With Encoding Variability*

Next, we wanted to explore a more realistic scenario in which the contribution of sampling variability to overall variability is small, relative to other forms of variability (like encoding variability) that affect studied items and corresponding lures in tandem. Our earlier simulations demonstrated that sampling variability is negligible in brain-sized networks (Figure 10). When other forms of variability — apart from sampling variability — predominate, the familiarity difference should be more reliable, and therefore cortex should benefit in a more robust fashion from use of an FC-corresponding-lure test.

To test this idea, we set out to increase the influence of encoding variability relative to sampling variability in the model. We added encoding variability to the model using the following simple manipulation: For each item

at study, the learning rate was scaled by a random number from the 0-to-1 uniform distribution. However, this manipulation by itself does not achieve the desired result; the influence of encoding variability is still too small relative to sampling variability, and overall performance levels with added encoding variability are unacceptably low. To boost the relative impact of encoding variability (and overall performance), we also increased the learning rate to three times its usual value, from .004 to .012. Under this regime, random scaling of the learning rate at study has a much larger effect on studied-item (and related-lure) familiarity than random differences in how well features are sampled. We should note that using a large learning rate has some undesirable side-effects (e.g., increased interference), but these side-effects are orthogonal to the questions we are asking here.

Figure 15a shows the cortical results, together with results from the hippocampal model that are discussed next. The corresponding vs. non-corresponding difference for the cortical model is much larger in these simulations than in the previous ones (Figure 14). Computing the average covariance/variance ratio for the four highest overlap conditions shows that  $R = .63$  for corresponding lures vs.  $R = -.06$  for non-corresponding lures. This is nearly double the covariance in the basic model (.63 vs. .31), and confirms our intuition that decreasing the contribution of sampling variability relative to encoding variability should increase covariance and boost performance in the FC corresponding condition.

#### *The Hippocampal Model*

Figure 15a also shows how use of corresponding vs. non-corresponding lures affects FC recognition in the hippocampal model. Like the cortical-model simulations, these hippocampal simulations incorporated encoding variability and they used a high learning rate (3X the normal value). To maximize hippocampal recognition performance, we used a recall-to-reject FC decision rule — if one item triggered mismatching recall, but the other item did not, we selected the second item; otherwise, we selected the item triggering a higher *match – mismatch* recall score. We also ran YN recognition simulations in both models using the same parameters (Figure 15b), so we could compare FC performance to YN performance. The YN hippocampal simulations also used a recall-to-reject rule.

The most interesting result is that FC corresponding and non-corresponding performance are almost identical in the hippocampal model. It seems clear that the same arguments about covariance benefiting FC-corresponding performance should hold for hippocampus as well as for cortex. Why then does the hippocampus behave differently than the cortex in this situation? This can be explained by looking at what happens on

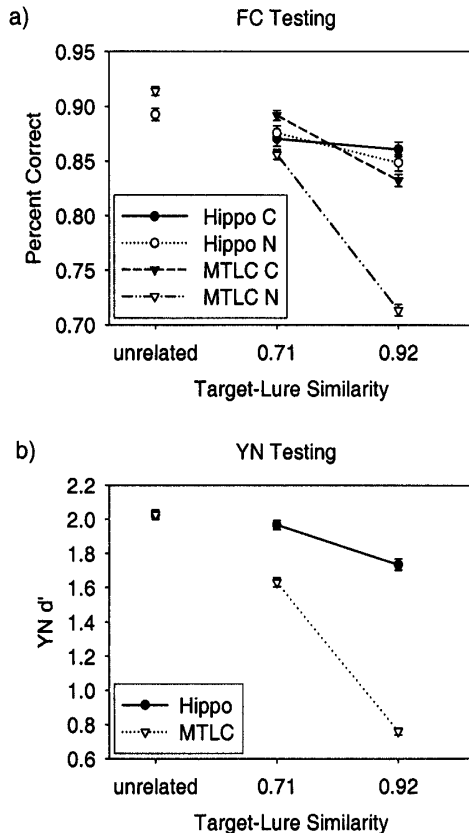


Figure 15: Cortical and hippocampal related-lure simulations incorporating strong encoding variability. (a) Shows cortical and hippocampal performance on FC-corresponding (C) and FC-non-corresponding (N) tests. Cortex performs better with corresponding vs. non-corresponding lures, but the hippocampus (using recall-to-reject) performs equally well with corresponding vs. non-corresponding lures. As a result, cortex is more strongly impaired (relative to the hippocampus) on FC-non-corresponding vs. FC-corresponding tests. (b) Shows cortical and hippocampal performance on YN tests. As in the previous related-lure simulations (Figure 12), cortex was severely impaired relative to the hippocampus on these tests.

trials where studied recall fails — on these trials, subjects can still respond correctly if the lure triggers *mis-matching* recall (and is rejected on this basis). The key insight is that studied recall and lure misrecall are *independent* when non-corresponding lures are used (in effect, subjects get two independent chances to make a correct response), but they are highly correlated when corresponding lures are used — if the studied item does not trigger any recall, the corresponding lure probably will not trigger any recall either. Thus, using corresponding lures can actually hurt performance in the hippocampal model by depriving subjects of an extra, independent chance to respond correctly (via recall-to-reject) on trials where studied recall fails. This harmful effect of covariance cancels out the beneficial effects of covariance described earlier. This analysis of the hippocampal system is supported by the fact that when we use the standard *match – mismatch* recall decision rule (which places less weight on mismatch than the recall-to-reject rule), the hippocampal model shows a substantial corresponding-lure advantage, just like cortex.

#### Testing the Model's Predictions

Figure 15 summarizes the key predictions that our models make regarding performance on YN and FC tests with related lures: The model predicts that cortex should perform worse than the hippocampus on YN tests with related lures, and on FC tests with non-corresponding related lures; but cortical performance should be relatively spared on FC tests with *corresponding* related lures. On FC tests with corresponding lures, the presence of covariance between studied items and related lures helps cortical performance, but it can actually harm hippocampal performance by depriving subjects of opportunities to benefit from recall-to-reject. These factors, taken together, work to push cortical and hippocampal performance closer together.

One way to test the model's predictions is to look at recognition in patients with focal, relatively complete hippocampal damage. Presumably, these patients are relying exclusively on MTLC familiarity when making recognition judgments (in contrast to controls, who have access to both hippocampal recall and MTLC familiarity). As such, patients should perform poorly relative to controls on tests where hippocampus outperforms cortex, and they should perform relatively well on tests where hippocampus and cortex are evenly matched. Applying this logic to the above results, patients should be impaired on YN recognition tests with related lures (because these tests load heavily on recall) but they should perform relatively well on FC-corresponding tests with related lures, and on tests with unrelated lures (regardless of test format).

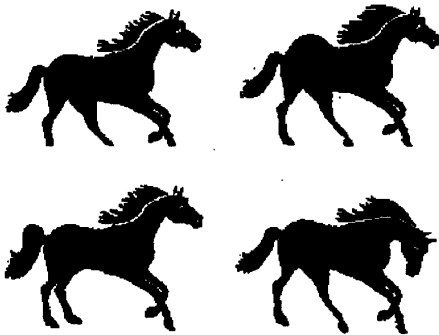


Figure 16: Sample stimuli from the Holdstock et al. (in press) related lure experiment. Participants studied pictures of objects (e.g., the horse shown in the upper left). At test, participants had to discriminate studied pictures from three highly related lures (e.g., the horses shown in the upper right, lower left, and lower right).

To test this prediction, we collaborated with Andrew Mayes and Juliet Holdstock to test patient YR, who suffered focal hippocampal damage, sparing surrounding MTLC regions (for details of the etiology and extent of YR's lesion, see Holdstock et al., in press). YR is severely impaired at recalling specific details — thus, YR has to rely almost exclusively on MTLC familiarity when making recognition judgments. Holdstock et al. (in press) developed YN and FC tests with highly related lures that were closely matched for difficulty, and administered these tests to patient YR and her controls. Figure 16 shows sample stimuli from this experiment. Results from this experiment can be compared to results from 15 other YN item recognition tests and 25 other FC item recognition tests that used lures that were less strongly related to studied items (Mayes et al., submitted); we will refer to these tests as the YN-low-relatedness and FC-low-relatedness tests, respectively.

Figure 17 shows that, exactly as we predicted, YR was significantly impaired on a YN recognition test that used highly related lures, but showed relatively spared performance on an FC version of the same test (YR actually performed slightly *better* than the control mean on this test). This pattern can not be explained in terms of difficulty confounds (i.e., YR performing worse, relative to controls, on the more difficult test) — controls found the YN test with highly related lures to be slightly easier than the FC test. Figure 17 also shows that YR was, on average, unimpaired on YN-low-relatedness and FC-low-relatedness tests. YR performed worse on the YN test with highly related lures than on any of the 15 YN-

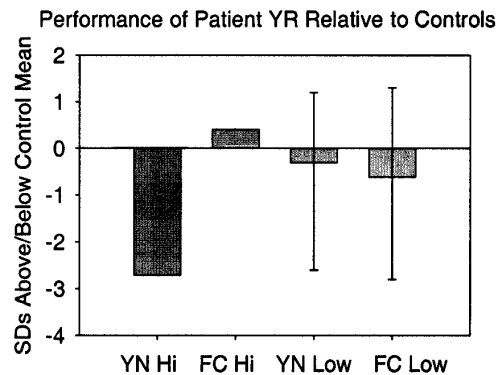


Figure 17: Performance of YR relative to controls on matched YN and FC-corresponding tests with highly related (Hi) lures; the graph also plots YR's average performance on 15 YN tests and 25 FC tests with less strongly related (Low) lures. YR's scores are plotted in terms of number of standard deviations above or below the control mean. For the YN and FC low-relatedness tests, error bars indicate the maximum and minimum z-scores achieved by YR (across the 15 YN tests and the 25 FC tests, respectively). YR was significantly impaired relative to controls on the YN test with highly related lures (i.e., her score was  $> 1.96$  SDs below the control mean) but YR performed slightly better than controls on the FC test with highly related lures. YR was not significantly impaired, on average, on the tests that used less strongly related lures.

low-relatedness tests; this difference can not be attributed to the YN-low-relatedness tests being easier than the YN test with highly related lures: YR showed unimpaired performance on the 8 YN low-relatedness tests that controls found to be *more difficult* than the YN test with highly related lures; for these eight tests her mean z-score was 0.04 ( $SD = 0.49$ ; minimum = -0.54; maximum = 0.65; J. Holdstock, personal communication). We have yet to test the model's prediction regarding use of FC-corresponding vs. FC-non-corresponding tests with related lures; based on the results shown in Figure 15, the model predicts that YR will be more strongly impaired on FC tests with non-corresponding (vs. corresponding) related lures.

### Simulation 5: Associative Recognition and Sensitivity to Conjunctions

In this section, we explore the two networks' sensitivity to feature conjunctions. The hippocampus' ability to rapidly encode and store feature conjunctions is not in dispute — this is a central feature of practically all theories of hippocampal functioning, including ours



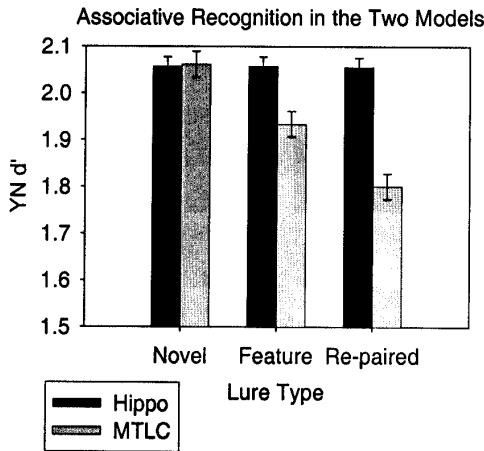


Figure 18: Results of YN associative recognition simulations in the cortical and hippocampal models. Using parameters that yield matched performance for unrelated (novel) lures, cortex is impaired relative to the hippocampus at associative recognition; importantly, cortex performs well above chance on the associative recognition tests.

(e.g., Rudy & Sutherland, 1995; Squire, 1992; Rolls, 1989; Teyler & Discenna, 1986). By contrast, many theorists have argued that neocortex is not capable of rapidly forming new conjunctive representations (i.e., representations that support differential responding to conjunctions vs. their constituent elements) on its own; see O'Reilly and Rudy (2001) for a review. To measure the two networks' sensitivity to conjunctions, we used an *associative recognition* paradigm, in which participants study pairs of stimuli (A-B, C-D); at test, subjects have to discriminate between studied pairs and associative lures generated by recombining studied pairs (A-D, B-C). To show above-chance associative recognition performance, a network must be sensitive to whether features occurred *together* at study — sensitivity to individual features does not help discriminate between studied pairs and recombined lures.

In our associative recognition simulations, 20 item pairs were presented at study — each “pair” consisted of a 12-slot pattern concatenated with another 12-slot pattern; at test, studied pairs were presented, along with three types of lures: associative (*re-paired*) lures, *feature* lures (generated by pairing a studied item with a non-studied item), and *novel* lures (generated by pairing two nonstudied items). The first set of associative recognition simulations used YN testing.

Figure 18 shows the results. It is clear that the hippocampus is unaffected by any of the lure manipulations

— its natural tendency to develop conjunctive representations of inputs means that it is not fooled by associative (re-paired) lures. In this simulation, associative lures did not trigger *any* recall. This is a consequence of the fact that we cued memory in this simulation with both pair items simultaneously; the presence of novel conjunctions in the test cue leads to pattern separation. As discussed below, an alternate approach to associative recognition is to cue with one pair item at a time; this approach (using partial cues) tends to trigger pattern completion as opposed to pattern separation (O'Reilly & McClelland, 1994).

There are two important conclusions to be gleaned from the cortical model results: First, the cortical model performs worse than the hippocampal model on the associative recognition test, indicating that MTLC familiarity is relatively *less* sensitive to conjunctions than hippocampal recall. However, cortical performance on the associative recognition test is well above chance — this indicates that cortex is sensitive (to some degree) to feature co-occurrence in addition to individual feature occurrence.

The ability of the cortical model to encode stimulus conjunctions can be explained in terms of the fact that cortex, like the hippocampus, uses sparse representations (as enforced by the k-winners-take-all algorithm). The kWTA algorithm forces units to compete to represent input patterns, and units that are sensitive to multiple features of a given input pattern (i.e., feature conjunctions) are more likely to win the competition than units that are only sensitive to single input features. Representations are more conjunctive in the hippocampus than in cortex because representations are more sparse (i.e., there is stronger inhibitory competition) in the hippocampus than in the cortex. Importantly, this explanation is in keeping with the idea that practically all differences between cortex and the hippocampus can be placed on a continuum — i.e., it is wrong to say that hippocampus is sensitive to conjunctions and cortex is not, or that hippocampus uses non-overlapping representations but cortex uses overlapping representations; in each of these cases, the difference is better described as a matter of degree (i.e., hippocampus is *more* sensitive to conjunctions, and hippocampal representations overlap *less* than cortical representations).

### Effects of Test Format

Although the model predicts that the hippocampus will outperform cortex on YN associative recognition tests, the model also predicts that the hippocampus will show less of an advantage on FC associative recognition tests where subjects have to choose between a studied item and an *overlapping* re-paired lure (i.e., study A-B,

C-D; test A-B vs. A-D). Typically these overlapping-lure tests are structured in a way that emphasizes the shared item: Subjects are asked, "which of these items was paired with A: B or D?". We argue that this format encourages subjects to adopt a *cued recall* strategy whereby they cue with the shared item (A).

When subjects cue with individual items (as opposed to item pairs), hippocampal associative recognition performance depends critically on recall-to-reject; single items from re-paired lures will frequently trigger recall, but when this happens the lure can typically be rejected based on mismatch (e.g., for the test pair A-D, subjects might cue with A and then recall that A was studied with B, not D). The hippocampal model performs relatively poorly on FC tests with overlapping choices (assuming that subjects cue with the shared item) because — in this situation — matching and mismatching recall are *completely redundant*: If A triggers recall of the A-B pair, subjects can respond correctly either by choosing B (based on match) or by rejecting D (based on mismatch); if A fails to trigger recall, subjects will not be able to respond correctly based on match or mismatch. In both cases, rejecting items based on mismatch confers no extra benefit above what you would get from accepting items based on match. In contrast, subjects do benefit from paying attention to mismatch on FC associative recognition tests with non-overlapping choices (i.e., study A-B, C-D, E-F; test A-B vs. C-F) because the probability of the lure triggering recall-to-reject is independent of the probability of the studied item triggering matching recall. Subjects also benefit from paying attention to mismatch on YN tests (when they cue with single items) insofar as this allows them to confidently reject re-paired lures.

Our next simulation is a simple demonstration of the fact that, in our model, FC tests with overlapping lures (*FC-OLAP* tests) prevent subjects from benefiting from recall-to-reject; it compares hippocampal associative recognition performance on FC-OLAP tests and FC tests with non-overlapping lures (*FC-NOLAP* tests). On both tests, we cued with the first pair item and measured how well recalled information matches the second pair item; on FC-OLAP tests, we cued with the shared pair item for both test alternatives; on FC-NOLAP tests we cued with the first item of each test alternative. We ran one version of the simulation where responses were based purely on matching recall, and we ran another version of the simulation where subjects used a recall-to-reject rule.

We had to adjust some model parameters to get the model to work well using partial cues; specifically, we used a higher-than-usual learning rate (.03 vs. .01) to help foster pattern completion of information not in the cue,

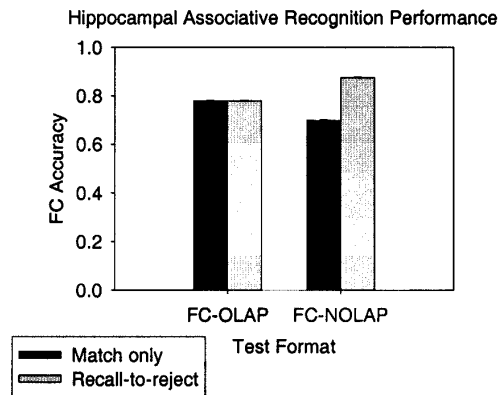


Figure 19: Associative recognition performance in the hippocampal model, as a function of recall decision rule (match only vs. recall-to-reject) and test format (FC-OLAP vs. FC-NOLAP). FC-NOLAP performance benefits from use of recall-to-reject, but FC-OLAP performance does not benefit at all. When only matching recall is used, OLAP performance is better than NOLAP performance, but when recall-to-reject is used, NOLAP performance is better than OLAP performance.

and we increased the activation threshold for counting a feature as recalled (from .90 to .95) to compensate for the fact that the output of the model is messier with partial cues.

Figure 19 shows the results of this simulation. FC-NOLAP performance benefits strongly from use of recall-to-reject (vs. the "match only" rule), but FC-OLAP performance does not benefit at all. When only matching recall is used, OLAP performance is better than NOLAP performance; this reflects the beneficial effects of covariance in recall scores triggered by the shared cue (i.e., when you cue with the shared item, recalled information always matches the studied item at least as much as the lure, usually more). In contrast, when recall-to-reject is used, NOLAP performance is better than OLAP performance. Consistent with this prediction, Clark, Hori, and Callan (1993) found better performance on an FC-NOLAP associative recognition test than on an FC-OLAP associative recognition test. They explained this finding in a manner that is consistent with our account — they argued that subjects were using recall of studied pairs to reject lures, and that subjects have more unique (independent) chances to recall useful information in the NOLAP condition.

### Tests of the Model's Predictions

In summary, the model predicts that hippocampus will outperform cortex on YN associative recognition

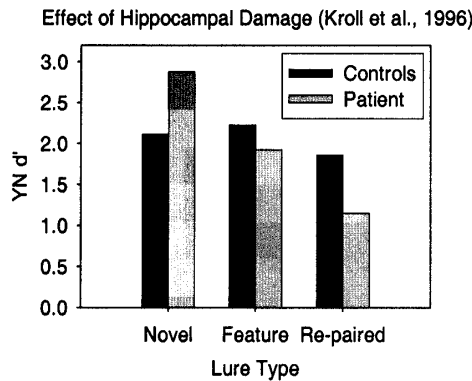


Figure 20: Results of Experiment 1 from Kroll et al. (1996), which examined associative recognition in a patient with bilateral hippocampal damage;  $d'$  scores for the patient and controls were computed based on average hit and false alarm rates published in Table 3 of that paper. The patient performed better than adult controls at discriminating studied items from novel lures but was worse than controls at discriminating studied items from feature lures (where one part of the stimulus was old and one part was new) and was much worse than controls when lures were generated by re-combining studied stimuli. The pattern reported here is qualitatively consistent with the model's predictions as shown in Figure 18.

tests for two reasons: First, when subjects cue with both parts of a re-paired lure, this typically does not result in *any* recall because of hippocampal pattern separation; second, when subjects cue with single items, the hippocampal model can typically reject re-paired lures based on mismatching recall. Use of an FC-OLAP test (where subjects cue with the shared item) reduces the hippocampal advantage because use of single-item cues deprives subjects of benefits conferred by hippocampal pattern separation, and cuing with the shared item prevents subjects from benefiting from recall-to-reject. The implications of these simulation results for patient performance are clear: Patients with focal hippocampal lesions should be impaired, relative to controls, on YN associative recognition tests, but they should be relatively less impaired on FC-OLAP associative recognition tests.

No one has yet conducted a direct comparison of how well patients with hippocampal damage perform relative to controls, as a function of test format. However, there are several relevant data points in the literature. The only published study looking at YN associative recognition in patients with focal hippocampal damage was conducted by Kroll et al. (1996). Results from Experiment 1 of Kroll et al. (1996) are plotted in Figure 20; in this experiment, participants studied 2-syllable words (e.g., "barter", "valley") and had to discriminate between stud-

List Type	Target Items			Interference Items		
short:	bike	robot	apple			
long:	bike	robot	apple	cat	tree	towel
weak interf.:	bike	robot	apple	cat	tree	towel
strong interf.:	bike	robot	apple	cat	tree	towel
				cat	tree	towel

Table 2: Interference conditions: list length compares short and long lists, list strength compares weak interference items with strong ones.

ied words and words created by re-combining studied words (e.g., "barley"). In keeping with the model's predictions, YN associative recognition performance was impaired in a patient with bilateral hippocampal damage (caused by anoxia) but YN discrimination with novel lures (where neither part of the stimulus was studied) was intact; furthermore, even though the patient was impaired at associative recognition, the patient's performance in this experiment was above chance. This is consistent with the idea that cortex is sensitive (to some degree) to feature conjunctions. However, this study does not speak to whether cortex can form *novel* associations between previously unrelated stimuli — because stimuli (including lures) were familiar words, subjects do not necessarily have to form a *new* conjunctive representation to solve this task.

Two studies (Vargha-Khadem et al., 1997; Mayes et al., 2001) have examined how well patients with focal hippocampal damage perform on FC-OLAP tests where subjects are cued with one pair item and have to say which of two items was paired with that item at study. The Vargha-Khadem et al. (1997) study used unrelated word pairs, nonword pairs, familiar face pairs, and unfamiliar face pairs as stimuli, and the Mayes et al. (2001) study used unrelated word pairs as stimuli. In all of these studies, the hippocampally-lesioned patients were unimpaired. This is consistent with the model's prediction that patients should perform relatively well, compared to controls, on FC-OLAP tests. Furthermore, the patients' excellent performance on these tests despite having large hippocampal lesions, coupled with the fact that the tests used novel pairings, provides clear evidence that cortex is capable of forming new conjunctive representations (that are strong enough to support recognition, if not recall) after a single study exposure.

### Simulation 6: Interference and List Strength

We now turn to the fundamental question of how interference affects recognition: How does studying an item affect recognition of other (previously studied)

items? This question has been studied empirically in the context of the *list length* manipulations, which involve adding new items to the study list, and *list strength* manipulations, which involve strengthening memory for some, but not all, list items (Table 2). The relationship between these paradigms can be made clear with a simple example. First, we construct a list of *target items* (e.g., BIKE, ROBOT, APPLE). In a list length manipulation, we compare recognition of these target items by themselves (a *short list*) to recognition of target items after we add *interference items* to the list (e.g., CAT, TREE, TOWEL — the *long list*). If studying these additional items interferes with memory for the target items, performance should decrease in the long list condition. In a list strength manipulation, we compare recognition of target items in a list where targets and interference items are presented once (the *weak interference list*) with recognition of target items in a list where targets are presented once, and interference items are presented multiple times (the *strong interference list*). Thus, list strength effects measure how much *repeated study* of other items interferes with memory for the target items. Put another way, list length measures how much studying other things once interferes, while list strength measures additional interference from repeated study of these items.

What makes these interference paradigms particularly interesting in the context of recognition memory is that there appears to be a dissociation between list length and list strength effects: list length effects are reliably, though not universally, observed (e.g., Ohrt & Gronlund, 1999; Murnane & Shiffrin, 1991a; Gillund & Shiffrin, 1984; but see Dennis & Humphreys, 2001 for discussion of confounds present in some list length studies); by contrast, list strength effects appear to be non-existent or even sometimes slightly *negative* (i.e., recognition in the strong interference condition is actually slightly better than recognition in the weak interference condition) (Ratcliff, Clark, & Shiffrin, 1990). In other words, studying additional items once seems to cause interference, but studying them multiple additional times does not. Understanding why this can happen presents a challenge to any memory model, especially given that most models have a strong tendency to produce interference effects. In particular, it has been argued that interference is inevitable in neural network models that use overlapping representations (e.g., Ratcliff, 1990; McCloskey & Cohen, 1989), leading some researchers to dismiss such models on the grounds that they can not explain null interference effects that have been reported in the literature (see, e.g., Murnane & Shiffrin, 1991a).

In this section, we first review how and why neural network models that use Hebbian learning are susceptible to interference. Then, we explore the cortical and hippocampal models' susceptibility to interference, starting

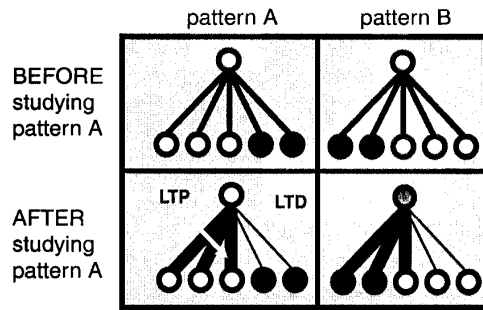


Figure 21: Two input patterns (A and B) that activate the same hidden unit (lighter colors = more activation). Studying pattern A increases weights to features that are *shared* by A and B (Hebbian LTP), and decrements weights to features that are *unique* to B (Hebbian LTD).

with the list strength paradigm. Our list strength simulations demonstrate that, although both networks are susceptible to interference (in the sense that new learning degrades stored traces), there are important differences in how interference manipulations affect recognition performance in the two networks. In particular, we show that the cortical network exhibits a null list strength effect on recognition sensitivity under many conditions, while the hippocampal network is more prone to showing list strength effects. The models make several novel, testable predictions about when interference effects should be obtained in recognition memory experiments — we summarize the results of several experiments conducted by Norman (submitted) that validate the models' predictions. After discussing list strength, we discuss the list length effect and explain how list length and list strength can have differential effects on recognition performance.

### General Principles of Interference in our Models

At the most general level, interference occurs in neural networks whenever a weight connecting two units is used in conflicting ways for two different memories. The level of overlap between different memory representations determines the extent to which weights are shared, and thus the potential for interference. This is why sparse representations in the hippocampus help to minimize interference — they minimize pattern overlap.

It is important to emphasize, however, that even if weights participate in storing multiple different memories, interference only arises if the weights are used in conflicting ways. In other words, interference does not occur for weights that encode shared features between two memories; rather, interference occurs for weights as-

sociated with discriminative (non-shared) features of the respective memories. Figure 21 illustrates this fact; it shows a simple network with a single hidden unit that receives input from five input units; initially, the hidden unit is activated above threshold by two different input patterns, A and B. The Hebbian learning rule used in our models dictates that weights to active input features get strengthened (LTP), and weights to inactive input features get weakened (LTD). Thus, when pattern A is studied, weights to features shared by patterns A and B increase due to Hebbian LTP (producing a stronger representation of these features), but weights to features that are unique to pattern B decrease due to Hebbian LTD. In short, studying pattern A degrades the network's representation of the discriminative features of pattern B (and vice-versa).

In the long run, this weakening of discriminative features is bad for recognition performance. If you train the network on a large number of overlapping patterns (e.g., several pictures of fish), the network will become more and more sensitive to features that are shared across the entire item set (e.g., the fact that all studied stimuli have fins), and it will become less and less responsive to the discriminative features of individual stimuli (e.g., the fact that one fish has a large green striped dorsal fin). However, these discriminative features are the only way to distinguish a studied fish item from a lure item that is also a fish, so weaker representations thereof can clearly lead to impaired performance. Whether or not recognition performance is actually harmed by this effect depends on the extent to which interference differentially affects responding to studied items and lures. In the next section, we explore this issue in the context of our two models.

### List Strength Results

We begin our exploration of interference by simulating the list strength paradigm as diagrammed in Table 2. The only difference is that — instead of strengthening items by presenting them repeatedly — we strengthened interference items by increasing the learning rate for these items (from .01 to .02). This approach to strengthening allows us to implement an arbitrarily powerful strengthening manipulation (by boosting the learning rate) without increasing the amount of time it takes to run the simulations.<sup>3</sup> In these simulations, the study list was comprised of 10 target items, followed by 10 in-

<sup>3</sup>In our models, strengthening by repetition and strengthening by increasing the learning rate have qualitatively similar effects; however, quantitatively, repetition has a larger effect on weights (e.g., doubling the number of presentations leads to more weight change than doubling the learning rate), because the initial study presentation alters how active units will be on the next presentation, and greater activity leads to greater learning (according to the Hebb rule).

terference items. We also manipulated average between-item overlap (ranging from 10% to 50%) to see how this factor interacts with list strength — intuitively, increasing overlap should increase interference.

Figure 22 shows the effect of list strength on recognition sensitivity in the two models, as a function of input overlap. In the cortical network (Figure 22a), there was no effect of list strength on recognition when input pattern overlap was relatively low (up to .26), but the list strength effect (*LSE*) was significant for higher levels of input overlap. In contrast, the hippocampal network showed a significant *LSE* for all levels of input overlap (Figure 22b); the size of the hippocampal *LSE* increased with increasing overlap (except in the .5 overlap condition, where the *LSE* was compressed by floor effects). Figure 22c directly compares the size of the *LSE* in the two models.

We also measured the *direct* effect of strengthening interference items on memory for those items; both models exhibited a robust *item strength* effect whereby memory for interference items was better in the strong interference condition (e.g., for 20% input overlap, interference-item  $d'$  increased from 2.13 to 3.22 in the hippocampal model; in the cortical model,  $d'$  increased from 2.08 to 3.12), thereby confirming that our strengthening manipulation was effective.

The data are puzzling: For moderate amounts of overlap, the hippocampus shows an interference effect despite its ability to carry out pattern separation, and cortex — which has higher baseline levels of pattern overlap — does not show an interference effect. We address the hippocampal results first.

### Interference in the Hippocampal Model

The hippocampal list strength effect is, at a general level, easy to understand because it basically amounts to a neural network exhibiting interference effects (as they are prone to do). At a more detailed level, there are two key points. First, even though there is less overlap between representations in the hippocampus than in cortex, there is still some overlap (as shown in our pattern separation simulations; see Figure 6). These overlapping units cause interference — specifically, recall of discriminative features of studied items is impaired through Hebbian LTD. Second, recall of discriminative features of lures (i.e., features of lures not shared by studied items) is at floor, because of the thresholded nature of hippocampal recall — a feature will only be recalled if weights to that feature were strengthened at study. Because recall of discriminative features of lures is at floor, it can not decrease as a function of interference. Putting these two points together, the net effect of interference is to move the studied distribution downwards towards the

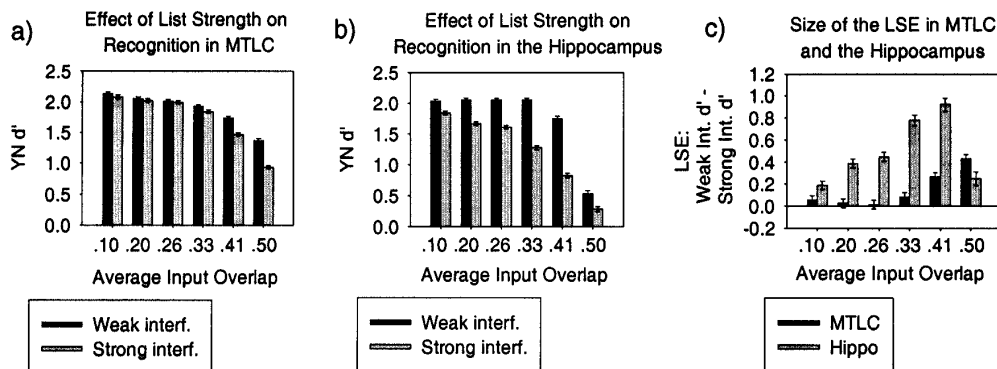


Figure 22: Results of list strength simulations in the two models. The study list consisted of 10 target items, followed by 10 interference items; list strength (*weak interference* vs. *strong interference*) was manipulated by doubling the interference-item learning rate in the strong interference condition. (a) shows MTLC results; (b) shows hippocampal results; (c) re-plots data from (a) and (b) as list strength difference scores (weak interference  $d'$  - strong interference  $d'$ ) to facilitate comparison across models. For low-to-moderate levels of overlap (up to .26), there was a significant LSE in the hippocampal model but not in the cortical model; for higher levels of overlap there was an LSE in both models.

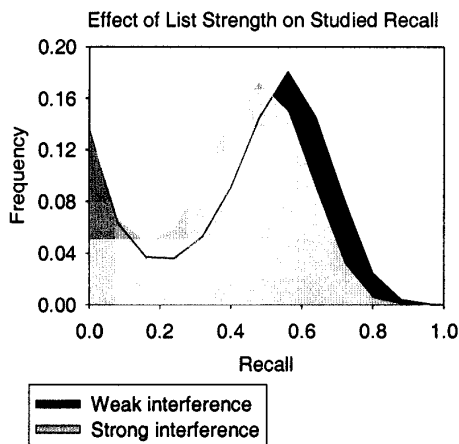


Figure 23: Studied recall histograms for the strong and weak interference conditions, 20% overlap condition. Increasing list strength pushes the studied recall distribution to the left (towards zero).

at-floor lure distribution, which increases the overlap between distributions and therefore impairs discriminability (Figure 23). We should note that, in high-overlap situations, increasing list strength may lead to increased recall of shared prototypical features (insofar as weights to these features get strengthened). However, this effect is not differential — it occurs equally for studied items and lures; therefore, it does not affect discriminability.

The above explanation implies that there will always be interference effects on hippocampal recall, insofar as there will always be some overlap, and overlap always leads to decreased recall of discriminative features of studied items. However, when ceiling effects are present for recall of studied stimuli, and when overlap between items is relatively low, interference effects may be very small and hard to detect. That is: When the studied recall distribution is located far to the right of the recall threshold, interference effects (which push this distribution to the left) may not lead to an appreciable decrease in above-threshold recall. We ran a list strength simulation in the hippocampal model using distinctive inputs (10% overlap) and a high learning rate for target items (.03, instead of .01) that produces strong recall. In the weak interference condition, 99.8% of items triggered above-threshold (i.e.,  $> .40$ ) recall, and 94% of items triggered perfect recall scores. Increasing list strength (by tripling the learning rate for interference items) led to a decrease in the proportion of items triggering perfect recall scores (91%) but the proportion of studied items triggering above-threshold recall scores was virtually unchanged (99.7%). Finally, we should note that — in experiments where encoding is not tightly controlled — ceiling effects might be present for recall of individual items even if overall recall performance is not at ceiling; to the extent that this occurs, it will curtail the effect of interference on recall (for more discussion of this issue, see Norman, submitted).

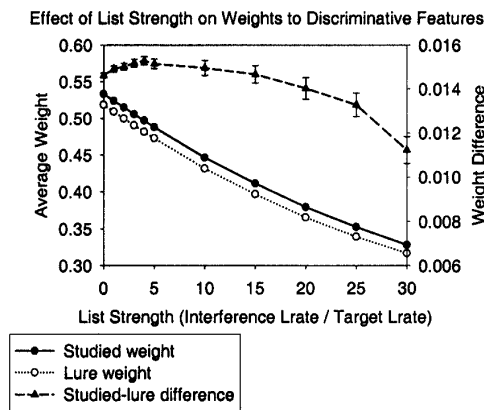


Figure 24: Average weights to discriminative features of target items and lures, as a function of list strength; the graph also plots the studied-lure weight difference. Overall, increasing list strength results in lower weights to discriminative features; there is also an interaction with item type, whereby weights to the discriminative features of lures decrease more quickly than weights to the discriminative features of studied target items; as a result, the target-lure gap in sensitivity to discriminative features actually increases (at first) with increasing list strength.

### Interference in the Cortical Model

Next, we need to explain why an LSE was not obtained in the cortical model (for low-to-moderate levels of input overlap). The critical difference between the cortical and hippocampal models is that lure familiarity is not at floor in the cortical network, thereby opening up the possibility that lure familiarity (as well as studied familiarity) might decrease as a function of interference. Discriminability is a function of the *difference* in studied and lure familiarity (as well as the variance of these distributions); therefore, if lure familiarity decreases as much as (or more) than studied familiarity as a function of interference, overall discriminability may be unaffected. This is in fact what occurs in the cortical model.

As discussed earlier, recognition performance depends critically on sensitivity to discriminative features of studied items and lures (sensitivity to shared features does not benefit recognition performance, because these features are equally present in studied items and lures). More specifically, the fact that studied items are more familiar than lures can be traced back to the fact that the network has stronger weights to the discriminative features of studied items than to the discriminative features of lures. Because sensitivity to discriminative features is such an important determinant of recognition performance, we developed a direct measure of this factor — we can then see how sensitivity to discriminative features

of studied items and lures (and, critically, the *difference* in sensitivity) changes as a function of interference. Sensitivity was measured as follows: For each target and lure item, we computed the average weight value linking the discriminative (i.e., non-prototypical) features of that item to that item's MTLC representation, prior to interference items being presented. Then, we measured the exact same weights after interference items were presented. List strength was manipulated by varying the learning rate for interference items.

Figure 24 plots the results of these simulations. Because of Hebbian LTD effects, increased learning of interference items reduces sensitivity to the discriminative features of both studied items and lures in a monotonic fashion. However, discriminability depends on the *difference* in sensitivity to features of studied items and lures — this difference initially rises slightly and then decreases with increasing interference. This rise happens because sensitivity to the discriminative features of lures decreases more rapidly than sensitivity to the discriminative features of studied items. This is the key to explaining the null (and sometimes trending negative) list strength effect in neocortex, and we will discuss why this occurs in the *Differentiation* subsection below. Of course, changes in the “target-lure sensitivity gap” must be weighed together with changes in variability, which can degrade performance even if the target-lure sensitivity gap increases. In the simulations reported here, variability does not increase enough to negate the aforementioned increase in the sensitivity gap; however, we can not rule out the possibility that adding other forms of variability to the model (and eliminating sampling variability; see the *Variability and Scaling Effects* section) might alter these predictions.

With enough interference, the target-lure sensitivity gap starts to shrink again. This is primarily attributable to floor effects on weights to discriminative features of lures, which tend to be smaller to begin with and thus have less to lose. This “floor effect” dynamic mirrors what occurs in the hippocampal model. Also, as interference increases, MTLC units that are predominantly sensitive to discriminative features start to *drop out* of item representations (i.e., because their weights are so weak, they no longer win the competition to be active) — these units are replaced by MTLC units that are relatively more sensitive to shared features. This dropout factor accelerates the inevitable shrinkage of the target-lure sensitivity gap. Figure 24 misses these dropout effects because it plots weights to MTLC units that were active prior to interference, regardless of whether they were still active after interference items were presented. Figure 25 is identical to Figure 24, except it only plots weights to MTLC units that were active *after* interference items were presented — therefore, the weights plot-

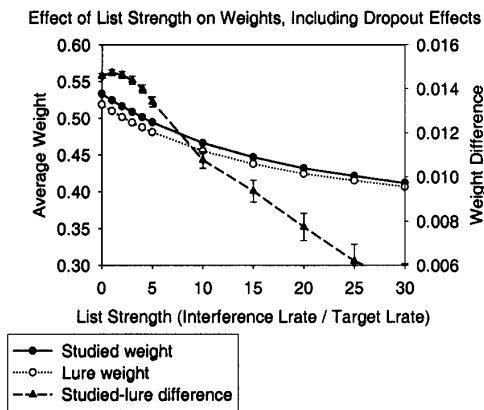


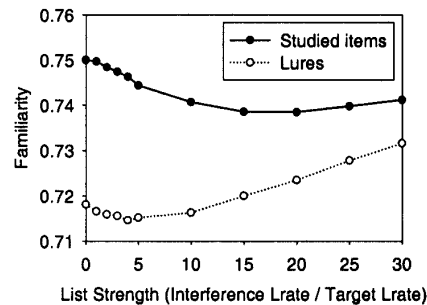
Figure 25: Figure 24, adjusted to reflect "dropout effects" (i.e., the fact that — with increasing interference — units sensitive to discriminative features tend to drop out of item representations; these units are replaced by other units that are less sensitive to discriminative features, and more sensitive to shared features). Note how the target-lure sensitivity gap decreases more rapidly in this figure than in Figure 24.

ted here reflect dropout effects in addition to the factors discussed earlier. When we factor in dropout effects, the target-lure sensitivity gap shrinks more rapidly as a function of interference.

To this point, we have focused on discriminative weights because they are the primary determinant of recognition performance. Interesting dynamics also occur in the raw familiarity scores triggered by targets and lures as a function of interference (Figure 26a). Initially, target and lure familiarity decline — the decrease in weights to discriminative features outweighs the concomitant increase in weights to shared features. However, with enough interference, target and lure familiarity both start to increase. This increase happens because weights to discriminative features approach floor (for both studied items and lures); therefore, these weights no longer decrease enough to offset the increase in weights to shared features. Because weights to discriminative features of lures hit floor before weights to discriminative features of studied items, the "upturn" in lure familiarity occurs before the upturn in studied-item familiarity.

Figure 26b plots the difference in target and lure familiarity scores. Initially, the difference increases slightly as a function of interference, but then the difference starts to shrink. These results are qualitatively identical to the results from our analytic weight simulations (Figure 25) and therefore validate our use of the "target-lure weight gap" as an analytic means of predicting the familiarity gap in the model.

a) Effect of List Strength on Raw Familiarity, 20% Overlap



b) Effect of List Strength on the Studied - Lure Familiarity Gap

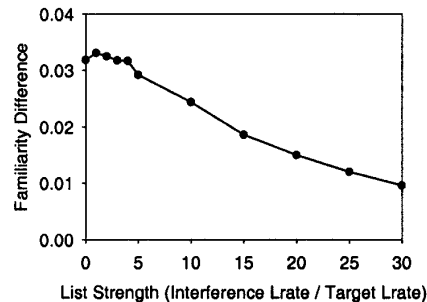


Figure 26: (a) Plot of how target and lure familiarity are affected by list strength (with 20% input overlap). Initially, target and lure familiarity decrease; however, with enough interference, target and lure familiarity start to increase. (b) Plot of the difference in target and lure familiarity, as a function of list strength; initially, the difference increases slightly, but then it decreases.

#### Boundary Conditions on the Null LSE

It should be clear from the above explanation that we do not always expect a null list strength effect in the cortical model. With enough interference the model's overall sensitivity to discriminative features always approaches floor and the studied and lure familiarity distributions converge. The amount of overlap between items determines how quickly the network arrives at this degenerate state — more overlap yields faster degeneration. When overlap is high, raw familiarity scores increase (and the familiarity gap decreases) right from the start; this is illustrated in Figure 27, which plots target and lure familiarity as a function of list strength, for 40.5% input overlap.

#### Differentiation

We have documented that lure representations initially degrade faster than studied representations, but we



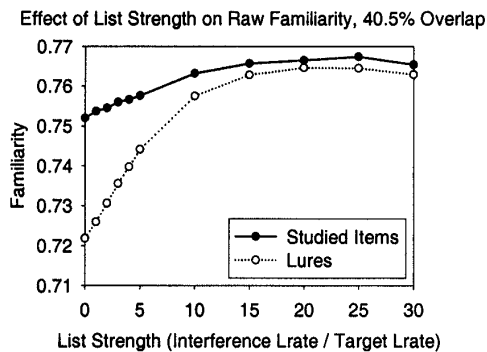


Figure 27: Plot of how target and lure familiarity are affected by list strength with 40.5% input overlap. When overlap is high, target and lure familiarity increase right from the start, and the target-lure familiarity gap monotonically decreases.

have not yet explained this result (which is the basis of the cortical null LSE). This finding can be explained in terms of the principle of *differentiation*, which was first articulated by Shiffrin, Ratcliff, and Clark (1990); see also McClelland and Chappell (1998). Shiffrin et al. argued that studying an item makes its memory representation more selective, such that the representation is less likely to be activated by other items. Intuitively, the more you know about an item, the less likely you are to confuse it with some other item.

In our model, differentiation is a simple consequence of Hebbian learning (as it is in McClelland & Chappell, 1998). The net effect of Hebbian learning in our model is to tune MTLC units so that they are more sensitive to the studied input pattern (because of LTP), and less sensitive to other, dissimilar input patterns (because of LTD). Because of this LTD effect, studied-item representations are *less likely to be activated by interference items* than lure-item representations; put another way, studied item representations will *overlap less* with the representations of interference items. As such, studied items suffer less interference than lures. As an example of how studying an item pulls its representation away from other items, in the 20% input overlap simulations the average amount of MTLC overlap between studied target items and interference items (expressed in terms of vector dot product) was .150, whereas the average overlap between lures and interference items was .154; this difference was highly significant.

### Summary and Predictions

In neural net models with Hebbian learning, interference necessarily degrades the network's responding to the discriminative (non-prototypical) features of studied

items and lures. Given the inevitability of degradation, the only way to avoid an interference effect on overall recognition discriminability is for responding to lures to degrade as much as (or more than) responding to studied items. This occurs in the cortical model because of differentiation: Studied items overlap less with interference items, therefore they suffer less interference than lures. This dynamic (whereby responding to lures initially degrades more than responding to studied items) does not apply to the hippocampus because of the thresholded nature of the hippocampal recall measure — lure recall is at floor, so it can not decrease with interference.

Thus, the main prediction from our models is that recognition based on hippocampal recall should generally exhibit an LSE, whereas recognition based on MTLC familiarity should not. Importantly, these patterns are not absolute, and are instead reliably affected by a number of experimental parameters. For example, strong target item encoding together with low input overlap can produce a ceiling effect on hippocampal recall that nullifies the hippocampal LSE. In the cortex, high levels of interference item strengthening produce list strength effects, as do high levels of input overlap. These stand as important testable predictions of the models.

### Testing the Model's Predictions

Consistent with the hippocampal model's prediction, some studies have found an LSE for *cued recall* (e.g., Kahana, submitted; Ratcliff et al., 1990) although not all studies that have looked for a cued recall LSE have found one (e.g., Bauml, 1997). However, all published studies that have looked for an LSE for recognition have failed to find one (Ratcliff et al., 1990; Murnane & Shiffrin, 1991a, 1991b; Ratcliff, Sheu, & Gronlund, 1992; Yonelinas, Hockley, & Murdock, 1992; Shiffrin, Huber, & Marinelli, 1995). Although this null LSE for recognition is consistent with our cortical model's predictions (viewed in isolation), it is nevertheless somewhat surprising that overall recognition scores do not reflect the hippocampal model's tendency to produce a recognition LSE.

One way to reconcile the null LSE for recognition with the model's predictions is to argue that hippocampal recall was not making enough of a contribution, relative to MTLC familiarity, on existing tests. This explanation leads to a clear prediction: List strength effects should be obtained for recognition tests and measures that load more heavily on the recall process. Norman (submitted) carried out three experiments to test this prediction, as summarized below.

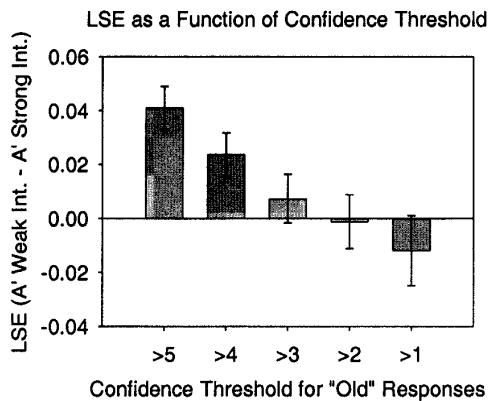


Figure 28: Plot of the size of the list strength effect, as a function of the confidence threshold used to compute "old" responses. When  $A'$  is computed using a high confidence threshold for saying "old" (confidence > 4, or confidence > 5), the LSE is significant. As the confidence threshold is lowered, the LSE gets smaller and eventually reverses direction.

#### Confidence Ratings

One way to test the model's prediction is to look at confidence rating data; several studies have found that recall is associated with high confidence "old" responses, whereas familiarity is associated with a range of confidence responses (Tulving, 1985; Yonelinas, in press; Yonelinas et al., 1996; Yonelinas, 1994). The high level of confidence associated with recall can be explained in terms of the *diagnosticity* (reliability) of the recall signal — the fact that false recall is rare means that, if you do recall an item, this is very strong evidence that the item was studied. The aforementioned results suggest that, if we compute recognition sensitivity based on high-confidence "old" responses (thereby isolating the contribution of recall), then we should find an LSE for recognition sensitivity.

In the experiment, Norman compared a weak interference condition (target and interference items studied once) with a strong interference condition (interference items studied six times, targets once). Participants studied concrete noun stimuli, using an encoding task ("would this item fit in a small box?"; 1.15 sec per word) that was designed to yield memory traces rich enough to support some recall, but not so distinctive as to yield ceiling effects on recall. At test, participants rated recognition confidence from 1 (sure new) to 6 (sure old); recognition sensitivity (indexed using  $A'$ ; Donaldson, 1993) was computed using different confidence thresholds for accepting an item as "old" (e.g., conf > 3 "old"; conf ≤ 3 "new").

Figure 28 shows the results of the experiment, plotting the size of the list strength effect ( $A'$  for targets in the weak interference condition minus  $A'$  for targets in the strong interference condition) as a function of the confidence threshold used to compute "old" responses. As predicted, a significant LSE on recognition sensitivity emerged when  $A'$  was computed using a high confidence threshold (4 or 5) for accepting an item as "old." This is the first time that anyone has documented a significant list strength effect on recognition sensitivity. When  $A'$  was computed using a lower confidence threshold, we obtained the same null (non-significant) list strength effect that other studies have found. Indeed, the results show a linear trend whereby the list strength effect decreases monotonically as the confidence threshold is lowered; this is consistent with the idea that lowering the confidence threshold for saying "old" increases the relative contribution of familiarity (thereby attenuating the list strength effect).

#### Self-Report Measures

A more direct way to isolate the contribution of recall is to look at self-report measures: Whenever a subject recognizes an item, you can simply ask them whether they recall studying the item or whether the item just seems familiar. Jacoby et al. (1997) showed that, if you make some assumptions about recall and familiarity (most prominently, independence), it is possible to use self-report data to separately examine how manipulations like list strength affect recall, and how they affect familiarity (the *independence remember-know*, or *IRK* procedure). Specifically, you can estimate  $P(R)$ , the probability of recalling a studied item, and  $Fd'$ , familiarity-based discrimination. The CLS model's prediction in this context is clear: List strength should affect the derived measure of recall, but not the derived measure of familiarity. Norman (submitted) tested this prediction using a paradigm that was structurally very similar to the paradigm used in the confidence-rating experiment, except — instead of giving confidence ratings — participants had to say whether the item was "old" or "new", and if they responded "old", participants had to say whether they *remembered* specific details (i.e., they recalled the item) or whether the item just seemed *familiar*.

Figure 29 shows the results of the experiment, plotting the size of the list strength effect for the derived measure of recall,  $P(R)$ , the derived measure of familiarity,  $Fd'$ , and for old/new recognition sensitivity (indexed using  $A'$ ). In this experiment, the effect of list strength on old/new recognition sensitivity was nonsignificant, replicating the null LSE obtained by Ratcliff et al. (1990). However, if you break recognition into its component processes, it is clear that list strength does affect per-

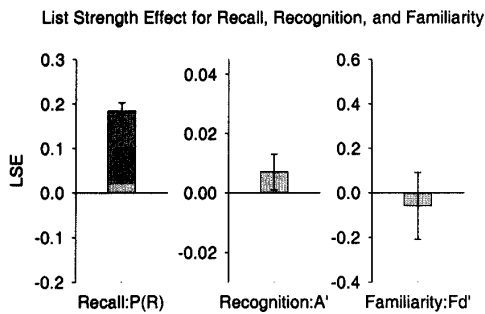


Figure 29: Plot of the size of the list strength effect for three dependent measures:  $P(R)$ , the derived measure of recall;  $A'$ , our measure of old/new recognition sensitivity; and  $Fd'$ , the derived measure of familiarity-based discrimination. The LSE was significant for recall, but the LSE was not significant for familiarity-based discrimination or old/new recognition sensitivity.

formance — as predicted, there was a significant LSE for recall; and there was a trend towards a *negative* LSE for familiarity-based discrimination. These results are of course contingent on the validity of the various assumptions that underlie the measurement procedure (e.g., independence, which we explore later), but they are highly consistent with the results of the first experiment.

#### Lure Relatedness

Yet another way to isolate the influence of recall is to use related lures at test — as discussed earlier, the model predicts that yes-no (YN) recognition tests with related lures should load heavily on recall, relative to tests with unrelated lures. Thus, we would expect a larger list strength effect on YN test with related lures, than on a YN test with unrelated lures. To test this hypothesis, Norman (submitted) used a plurals recognition paradigm (Hintzman et al., 1992; Curran, 2000) in which participants studied singular and plural words. At test, participants were instructed to say “old” if the test word exactly matched a studied word, and to say “new” otherwise; there were two kinds of lures: related switched-plurality (SP) lures (e.g., study “scorpion”, test “scorpions”) and unrelated lures. The model predicts that the ability to discriminate between studied words and related SP lures should depend on recall. Thus, we should find a significant list strength effect for studied vs. SP discrimination, but not necessarily for studied vs. unrelated discrimination, which can also be supported by familiarity.

Furthermore, we can also look at SP vs. unrelated *pseudodiscrimination*, i.e., how much more likely are subjects to say old to related vs. unrelated lures. Familiarity boosts pseudodiscrimination (insofar as SP lures

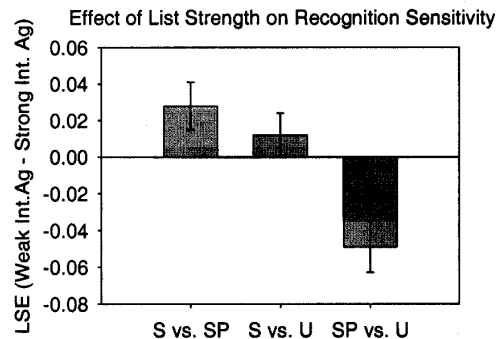


Figure 30: Results from the plurals LSE experiment. In this experiment, recognition sensitivity was measured using  $Ag$  (Macmillan & Creelman, 1991). The graph plots the size of the list strength effect for three different kinds of discrimination: Studied vs. related switched-plurality lures (S vs. SP); studied vs. unrelated lures (S vs. U); and related vs. unrelated lure pseudodiscrimination (SP vs. U). There was a significant LSE for studied vs. related lure discrimination, and there was a significant *negative* LSE for related vs. unrelated lure pseudodiscrimination.

will be more familiar than unrelated lures), but recall of plurality information lowers pseudodiscrimination (by allowing subjects to confidently reject SP lures). If list strength boosts familiarity-based discrimination, but lowers recall, both of these effects will work in concert to boost pseudodiscrimination. Hence, we predict a large *negative* LSE for pseudodiscrimination (i.e., it should be higher in the strong interference condition than the weak interference condition).

Apart from the plurality manipulation, the design of this experiment was very similar to the design of the prior two experiments.

The results from this experiment are shown in Figure 30. As predicted, the LSE for studied vs. SP lure discrimination was significant; the LSE for studied vs. unrelated lure discrimination was nonsignificant; and the LSE for SP lure vs. unrelated lure pseudodiscrimination was *negative* (and highly significant).

#### Summary

In summary, the models predict a dissociation whereby list strength effects should be obtained for recognition driven by hippocampal recall, but not for recognition driven by cortical familiarity. Norman (submitted) confirmed this prediction using three separate experiments, each of which used a different means of isolating the contribution of recall to recognition performance. However, the model's more complex list strength predictions (delineating boundary conditions on the null

cortical LSE, and the positive hippocampal LSE, as a function of factors like the amount of input pattern overlap) remain to be tested.

### Simulation 7: List Length and Dissociations with List Strength

Having explained how the CLS model can account for the null list strength effect for recognition, we now turn to the list length paradigm. As mentioned earlier, the memory literature indicates that there is a list length/list strength dissociation, whereby adding new items to the study list (increasing list length) hurts recognition, but additional study of these interfering items (increasing list strength) does not lead to further recognition deficits; for a particularly well-controlled demonstration of this dissociation, see Murnane and Shiffrin (1991a). Our models, in their most basic form, can not accommodate the length/strength dissociation — list length and list strength manipulations induce a comparable amount of interference; thus, the cortical model can not predict a list length effect at the same time that it predicts a null list strength effect.

However, our cortical model does produce the length/strength dissociation given the added postulate that the first presentation of an item leads to substantially more weight change than subsequent presentations of an item. As discussed earlier, learning about an item degrades the cortical network's ability to represent discriminative features of other items — the cortical network has a limited capacity to absorb these weight changes before discrimination starts to suffer (see Figure 25). If studying items for the first time (increasing list length) causes more weight change than repeating already-studied items (increasing list strength), then increasing list length will cause more degradation and therefore is more likely to push the network into the zone where it exhibits interference effects on  $d'$ .

This pattern of weight change is not an ad-hoc assumption to fit the data — it is directly supported by neurobiological research on long-term potentiation (LTP). Studies of LTP have found that presenting a potentiating stimulus generates a large increase in synaptic efficacy that decays to a smaller asymptotic value in a time window on the order of tens of minutes (e.g., Malenka & Nicoll, 1993; Bliss & Collingridge, 1993); additional stimulus presentations boost the asymptotic synaptic efficacy value but these repetition-induced adjustments tend to be small relative to the initial, transient increase in synaptic efficacy. To incorporate this dynamic in the model, we added transient *fast weights* that saturate after a single study trial, and decay rapidly (see Hinton & Plaut, 1987 for an early implementation of a similar

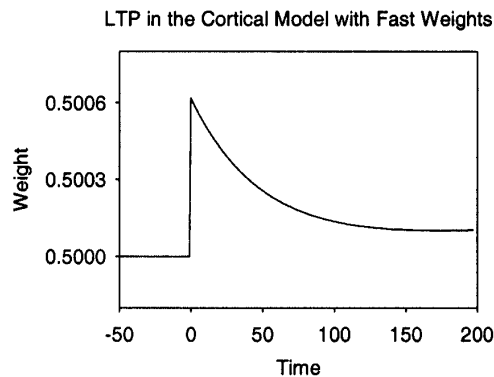


Figure 31: Illustration of how the weight between two units is affected by a potentiating stimulus (presented at time = 0) in a network with both fast weights and long-term weights. The size of the LTP effect is transiently large relative to its asymptotic value. The large, transient effect is attributable to fast weights, and the smaller, lasting effect is attributable to long-term weights. The pattern of weight change graphed here mirrors the dynamic observed in actual LTP studies (e.g., Nicoll et al., 1988; Malenka & Nicoll, 1993).

mechanism). This fast weight is added to the standard, slowly-adapting *long-term* weight to yield the effective strength of a given synapse; fast weights and long-term weights are adjusted separately by learning using our standard Hebbian LTP and LTD mechanisms, although the learning rate is higher for fast weights (see Appendix C for details). Figure 31 shows the results of a simple simulation in which we connected two units, activated them simultaneously, and measured how this affects the strength of the weight between the two units; because of the presence of fast weights, the resulting LTP effect is transiently large relative to its asymptotic value.

We ran closely matched list length and strength simulations using this fast weight mechanism in the cortical model. There were three different types of lists: a *weak/short* list comprised of 5 once-presented targets and 5 once-presented interference items; a *long* list comprised of 5 once-presented targets and 15 once-presented interference items; and a *strong* list comprised of 5 once-presented targets and 5 interference items presented 3 times. Note that the total number of interference presentations is the same in the *strong* vs. *long* conditions. Comparing memory for target items in the short/weak vs. long conditions provides a measure of the list length effect, because the number of interference items changes; comparing memory for target items in the short/weak vs. strong conditions provides a measure of the list strength effect, because the strength of each interference item changes, but the total number of interference items does

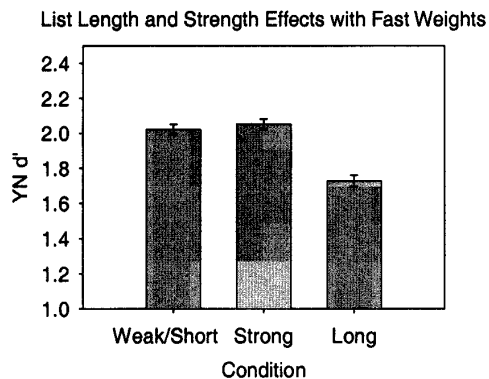


Figure 32: List length and strength effects in the cortical model with fast weights. These simulations used learning rate = .0025 for network weights; fast weight parameters are provided in Appendix C; there was 20% average overlap between input patterns. There was a highly significant list length effect ( $d'$  was lower in the *long* condition than in the *weak/short* condition) but no list strength effect ( $d'$  was almost identical in the *strong* and *weak/short* conditions).

not. To ensure that the lag between studying a target and testing that target was the same in all three conditions, we inserted 10 delay trials (in which no stimulus was presented) after the study list in the short/weak condition.

Figure 32 shows the simulation results. As in our previous list strength simulations (with 20% overlap), there was no list strength effect (i.e., target recognition was equivalent in the short/weak and strong conditions). However, there was a highly significant list length effect — target  $d'$  was worse in the long condition than in the short/weak condition. These simulations clearly demonstrate that adding rapidly-saturating, quickly-decaying fast weights to the cortical model results in a list length/list strength dissociation like the one obtained by Murman and Shiffrin (1991a). Importantly,  $d'$  for *interference* items was higher in the strong condition (2.73,  $SEM = .01$ ) than in the weak condition (2.38,  $SEM = .02$ ) — this shows that the null list strength effect observed here can not be attributed to use of an ineffective strengthening manipulation.

The model shows a list length effect because of the large learning rate associated with fast weights, which leads to large amounts of trace degradation — studying new interference items completely overwrites the fast weight values associated with discriminative features of other, overlapping items. The model does not show a list strength effect because fast weights saturate after a single study presentation; thus, subsequent presentations of interference items primarily affect long-term weights.

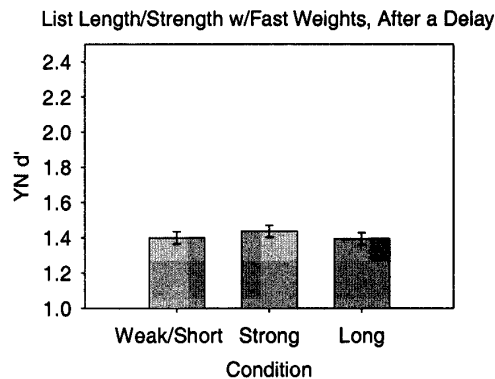


Figure 33: List length and strength effects in the cortical model with fast weights, after an 80-trial delay is interposed between study and test. In contrast to the (short-delay) simulations shown in Figure 32, which found a significant list length effect, the list length and list strength effects are both null after an 80-trial delay.

These long-term weights follow the exact same trajectory as a function of interference in this simulation as in our earlier simulations; increasing interference leads to a small decrease in weights to discriminative features of studied items, but this is offset by a similar decrease in weights to discriminative features of lures, so there is no interference effect on  $d'$ .

#### Decay and Delay Predictions

The fact that fast weights decay relatively quickly implies that the effects of fast weights should go away if a long delay is interposed between study and test. This means that the list length effect should disappear if fast weights are allowed to decay to zero, whereas any learning effects manifest in the permanent weights should persist. To test this prediction, we inserted 80 delay trials at the end of each study list in the length/strength simulation, thereby allowing the fast weights to decay to 20% of their maximal value (given the decay rate of .02).

In these “long-delay” simulations, there was no list strength effect or list length effect (Figure 33). However, item memory itself was only moderately diminished, and there was still a robust item strength effect ( $d'$  for interference items = 2.67 in the strong condition, vs. 1.53 in the weak condition). This clearly demonstrates that fast weights are responsible for the list length effect (and the length/strength dissociation), but item repetition effects are largely due to changes in long-term weights.

These simulation results lead to interesting predictions that can be tested in behavioral experiments. Specifically, we predict that interposing a delay between

study and test on the order of the transient-LTP decay constant (tens of minutes) should greatly diminish the list length effect. In support of this prediction, a recent study with a 5 minute delay between study and test did not show a list length effect (Dennis & Humphreys, 2001). The next step in testing this hypothesis will be to run experiments that parametrically manipulate study-test lag while measuring list length effects.

### Simulation 8: The Combined Model and the Independence Assumption

Up to this point, we have explored the properties of hippocampal recall and MTLC familiarity by presenting the same input patterns to separate hippocampal and neo-cortical networks — this approach is useful for analytically mapping out how the two networks respond to different inputs, but it does not allow us to explore interactions between the two networks. One important question that cannot be addressed using separate networks is the statistical relationship between recall and familiarity. As mentioned several times throughout this paper, all extant techniques for measuring the distinct contributions of recall and familiarity to recognition performance assume that they are stochastically independent (e.g., Jacoby et al., 1997). This assumption can not be directly tested using behavioral data because of the chicken-and-egg problems described in the Introduction.

To assess the validity of the independence assumption, we implemented a *combined model* in which the cortical system that computes familiarity serves as the input to the hippocampus — this arrangement more accurately reflects how the two systems are connected in the brain. The combined model is structurally identical to the hippocampal model except the projection from Input to EC.in has modifiable connections (and 25% random connectivity) instead of fixed 1-to-1 connectivity. Thus, the Input-to-EC.in part of the combined model has the same basic architecture and connectivity as the separate cortical model; this makes it possible to read out our *act.win* familiarity measure from the EC.in layer of the combined model.

There are, however, a few small differences between the cortical part of the combined model, and the separate cortical network. First, the EC.in layer of the combined model is constrained to learn “slotted” representations, where only one unit in each 10-unit slot is strongly active; limiting the range of possible EC representations makes it easier for the hippocampus to learn a stable mapping between CA1 representations and EC representations. Second, the EC.in layer for the combined model only has 240 units, compared to 1920 units in the MTLC layer of the separate cortical network. This reduced size

derives from computational necessity — use of a larger EC.in would require a larger CA1, which together would make the simulations run too slowly on current hardware. This smaller hidden layer in the combined model makes the familiarity signal more subject to sampling variability, and thus recognition  $d'$  is somewhat worse, but otherwise it functions just as before. We used the same basic cortical and hippocampal parameters as in our separate-network simulations, except we used input patterns with 32.5% overlap — this level of input overlap yields approximately 24% overlap between EC.in patterns at study.

As a first attempt at addressing the statistical relationship between the MTLC familiarity signal (*act.win*, read out from the EC.in layer) and the hippocampal recall signal (*match* – *mismatch* between EC.out and EC.in), we ran a simple recognition simulation and measured the correlation between these signals. A priori, one might think that, because the cortical and hippocampal systems are so closely interconnected, the two signals must be positively correlated to some extent. That is, sharper EC.in representations (which are associated with high familiarity scores) may propagate more efficiently into the hippocampus, leading to increased recall. Consistent with this view, we found a significant recall-familiarity correlation for studied items in this simulation ( $r = .27$ ,  $SEM = .02$ ). The recall-familiarity correlation for lures was not significantly different from zero ( $r = .02$ ,  $SEM = .02$ ) due to the fact that most runs of the network produced zero lure recall.

### Interference Induced Decorrelation

However, just because recall and familiarity are correlated (for studied items) in this simulation does not mean that they will always be correlated. In the next set of simulations, we show how the presence of interference between memory traces can reduce the recall-familiarity correlation. In the *Interference* section of the paper, we discussed how the two systems are differentially affected by interference: Hippocampal recall scores for studied items tend to decrease with interference; familiarity scores decrease less, and sometimes increase, because increased sensitivity to shared prototype features compensates for lost sensitivity to discriminative item-specific features. Insofar as items vary in how much interference they are subject to (due to random differences in between-item overlap), and interference pushes recall and familiarity in different directions, it should be possible to use interference as a “wedge” to decorrelate recall and familiarity.

We ran simulations measuring how the recall-familiarity correlation changed as a function of interference (operationalized using a list length manipulation).

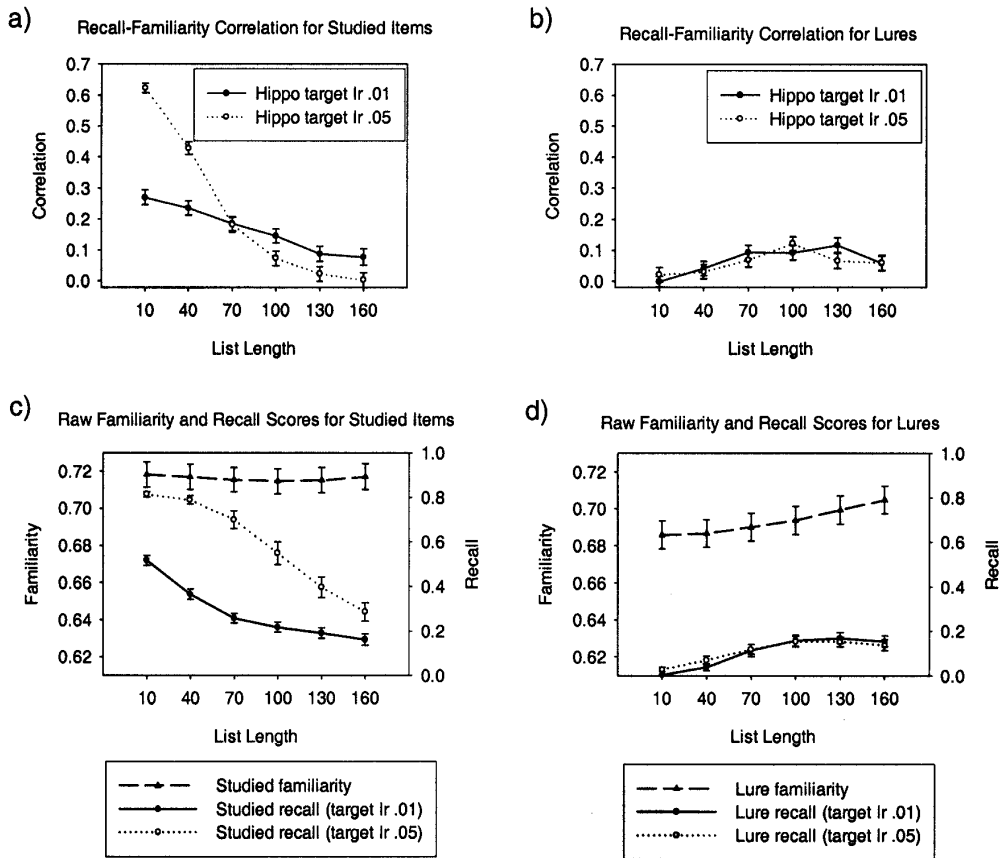


Figure 34: Simulations exploring how list length affects the recall-familiarity correlation for studied items and lures. The learning rate (lr) for target items in the hippocampus was manipulated (.01 vs. .05). Figures (a) and (b) plots the recall-familiarity correlation as a function of list length for studied items and lures, respectively. Increasing list length lowers the recall-familiarity correlation for studied items, but it leads to a slight increase in the recall-familiarity correlation for lures. Increasing list length leads to a sharper decrease in the studied-item recall-familiarity correlation when hippocampal target lr equals .05 vs. when target lr equals .01. (c) Plots how raw studied recall and studied familiarity scores are affected by list length. Recall decreases (more so for lr .05 than for lr .01) but familiarity stays relatively constant. (d) Plots how raw lure recall and lure familiarity scores are affected by list length. Both recall and familiarity increase slightly.

There were 10 target items followed by 0 to 150 (non-tested) interference items. We were concerned that floor-effects on item-specific recall might reduce the effectiveness of the list length manipulation; to address this concern, we manipulated the target item learning rate in the hippocampus, using both our standard value (.01) and a much larger value (.05) (interference-item learning rate was always .01). As expected, increasing list length lowers the recall-familiarity correlation for studied items (Figure 34a) in the model. There is also an interaction with hippocampal target learning rate (.01 vs. .05): With learning rate .05, the correlation starts out very high, but then decreases steeply and eventually goes to zero; with learning rate .01, the correlation starts out lower, decreases less steeply, and the curve asymptotes around correlation = .07.

We can get further insight into these results by looking at how interference affects raw familiarity and recall scores for studied items in these simulations (Figure 34c). When the hippocampal target learning rate is large (.05), recall starts out very high; basically, recall only fails in this condition when the cortical input is weak — this explains why the recall-familiarity correlation is initially very large. With increasing interference, recall decreases sharply but familiarity stays relatively constant; this differential effect of interference works to de-correlate the two signals. With a smaller hippocampal target learning rate (.01), recall decreases less sharply, because of floor effects. Once recall approaches floor, recall and familiarity are affected in a basically similar manner (i.e., not much); this lack of a differential effect explains why the recall-familiarity correlation does not continue to decrease all the way to zero.

Figure 34b shows that the recall-familiarity correlation for lures starts out at floor and then increases slightly as a function of interference. This can be explained by looking at raw recall and familiarity scores for lures (Figure 34d): With increasing interference, both lure familiarity and lure recall increase slightly; this is because increasing interference increases the cortical network's sensitivity to prototype features (thereby boosting familiarity), and it also increases the odds that the hippocampal network will recall these prototype features, instead of recalling nothing at all. This parallel effect of interference on familiarity and recall boosts the extent to which they are correlated.

### *Effects of Other Kinds of Variability*

Other sources of variability also affect the recall-familiarity correlation. For example, the high "baseline" recall-familiarity correlation exhibited by the model (in the absence of interference) is primarily due to sampling variability in EC.in; the small size of this layer (only 240

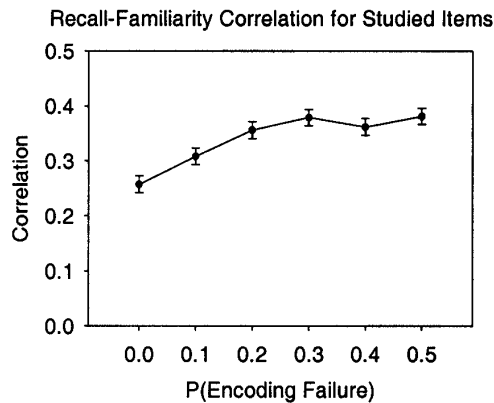


Figure 35: Simulations exploring the effect of encoding variability on the recall-familiarity correlation for studied items. As encoding variability (operationalized as the probability that subjects will "blink" and fail to encode half of an item's features at study) increases, the recall-familiarity correlation increases.

units) results in large, sampling-related fluctuations in the sharpness of cortical representations. These fluctuations in sharpness induce a correlation because they affect recall and familiarity in tandem — sharp representations trigger a large familiarity score and they propagate better into the hippocampus, bolstering recall. Therefore, reducing sampling variability (by increasing the size of the network) should reduce the recall-familiarity correlation.

Conversely, adding other forms of variability to the model that affect cortical representation strength should increase the recall-familiarity correlation. For example, at present, the model lacks encoding variability. Curran and Hintzman (1995) pointed out that encoding variability can boost the recall-familiarity correlation; if items vary substantially in how well they are encoded, poorly-encoded items will be unfamiliar and will not be recalled; well-encoded items will be more familiar, and more likely to trigger recall. We ran simulations in the combined model where we manipulated encoding variability by varying the probability of partial encoding failure (i.e., encoding only half of an item's features) from 0 to .50. The results of these simulations are presented in Figure 35; as expected, increasing encoding variability increased the recall-familiarity correlation in the model.

### *Summary and Implications*

In summary, the combined model gives us a principled means of predicting how different factors will affect the statistical relationship between recall and familiarity.



We identified two factors that push the recall-familiarity correlation in opposite directions: Interference makes the correlation smaller, and encoding variability makes the correlation larger. Importantly, even though the model predicts that recall and familiarity can be correlated, this does not mean that we have to abandon measurement techniques that assume independence. Instead, once we have identified the key determinants of independence using the model, we can use this information to guide the design of experiments so as to minimize violations of the independence assumption (e.g., by taking steps to minimize encoding variability, and to bolster interference).

### Simulation 9: Lesion Effects in the Combined Model

In this section, we show how the combined model can provide a more sophisticated understanding of the effects of different kinds of medial temporal lesions. One virtue of the combined model is that it lets us simulate how lesioning one structure affects performance in other structures — something that we could not do using our “separate networks” approach. Furthermore, the combined model lets us simulate how partial damage to various structures affects overall recognition performance. Whereas complete hippocampal lesions completely eliminate the contribution of recall, the effects of partial hippocampal lesions can be more complex; as we will show, partial hippocampal lesions can change the operating characteristics of hippocampal recall without eliminating the contribution of this signal to recognition performance.

### Lesion Controversies

As mentioned in the Introduction, studies examining the effects of hippocampal damage on recognition memory have obtained widely varying results; some studies have found deficits but others have found relatively spared performance. Baxter and Murray (2001b) recently conducted a meta-analysis of studies that have looked at hippocampal and perirhinal (MTLC) lesion effects on recognition in monkeys using a *delayed-nonmatching-to-sample* (DNMS) paradigm. They found, surprisingly, that partial hippocampal lesions may lead to *larger* recognition deficits than more complete lesions — in the meta-analysis, lesion size and recognition impairment were negatively correlated. In contrast, the Baxter and Murray meta-analysis found that perirhinal lesion size and recognition impairment were positively correlated. Thus, it may be possible to explain discrepant results across studies in terms of variability in the completeness of lesions, but in a somewhat counter-intuitive

way. However, Baxter and Murray’s claim is highly controversial; Zola and Squire (2001) re-analyzed the data in the Baxter and Murray meta-analysis, using a different set of statistical techniques that control, e.g., for differences in mean lesion size across studies, and found that the negative correlation between lesion size and impairment reported by Baxter and Murray (2001b) was no longer significant (although there was still a nonsignificant trend in this direction; see Baxter & Murray, 2001a for further discussion of this issue).

### Partial Lesion Simulations

An important missing piece to this puzzle is that no one (to date) has described a concrete mechanism that generates the hypothesized negative correlation between lesion size and impairment (although Baxter & Murray, 2001b outline several possible explanations in general terms). To explore this issue, we ran simulations mapping out the effects of partial vs. complete hippocampal and MTLC lesions in our combined model. The results of these simulations were consistent with the Baxter and Murray meta-analysis — partial hippocampal lesions (of a certain size) did impair recognition more than complete hippocampal lesions; by contrast, complete MTLC lesions impaired recognition more than partial MTLC lesions. Most importantly, these simulations provide a concrete and principled account of why partial lesions are especially harmful, and how more complete lesions ameliorate these harmful effects. The central principle is that partial hippocampal lesions impair the hippocampus’ ability to assign *pattern-separated* representations to stimuli; as a result, the amount of recall triggered by lures increases sharply — this becomes a source of noise that disrupts recognition performance, pulling it below the level that would be expected based on familiarity alone. Moving from a partial to a complete hippocampal lesion effectively removes this source of noise, thereby boosting recognition performance. Details of the lesion simulations are provided below.

In all of the lesion simulations, the size of the lesion (in terms of % of units removed) was varied from 0% to 95% in 5% increments. In the hippocampal lesion simulations, we lesioned all of the hippocampal subregions (DG, CA1, CA3) equally by percentage; in the MTLC lesion simulations, we lesioned EC.in. To establish comparable baseline (pre-lesion) recognition performance between the hippocampal and cortical networks, we boosted the cortical learning rate to .012 instead of .004; this increase compensates for the high amount of sampling variability present in the (240 unit) EC.in layer of the combined model.

In order to predict overall recognition performance when both processes are contributing, we had to make

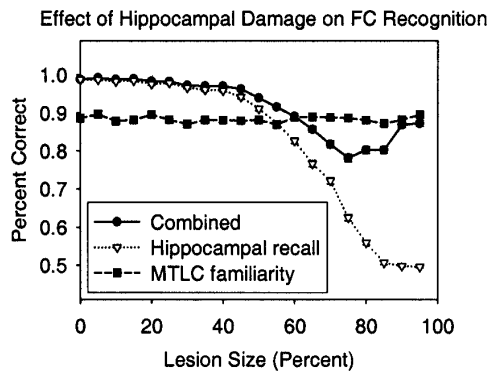


Figure 36: Effect of hippocampal damage on forced-choice (FC) recognition performance. This graph plots FC accuracy based on MTLC familiarity, hippocampal recall, and a combination of the two signals, as a function of hippocampal lesion size. FC accuracy based on MTLC familiarity is unaffected by hippocampal lesion size (insofar as the hippocampus comes after cortex in the processing chain). FC accuracy based on hippocampal recall declines steadily as lesion size increases. FC accuracy based on a *combination* of recall and familiarity is affected in a nonmonotonic fashion by lesion size: going from 0 to 75% hippocampal damage, combined FC accuracy declines; however, larger lesions lead to an *increase* in combined FC accuracy.

some assumptions about how subjects prioritize recall and familiarity information; specifically, we assume that recall takes precedence over familiarity. This assumption, which is shared by other dual-process theories (e.g., Jacoby et al., 1997) is reasonable in light of the fact that, in normal circumstances, recall is more diagnostic than familiarity. Furthermore, it is supported by the finding that recall is associated with higher average recognition confidence ratings than familiarity (e.g., Yonelinas, in press). In these simulations, we used a forced-choice test format to maximize comparability with the DNMS studies included in the Baxter and Murray (2001b) meta-analysis. Our decision rule was that if one item triggered a larger positive recall score (*match* – *mismatch*) than the other, then that item was selected. Otherwise, the decision falls back on familiarity. Certainly other rules would be possible, but this captures the critical functional properties in a simple way.

Figure 36 shows how hippocampal FC performance, cortical FC performance, and combined FC performance (using the decision rule just described) vary as a function of hippocampal lesion size. As one might expect, hippocampal FC performance decreases steadily as a function of hippocampal lesion size, while cortical performance is unaffected by hippocampal damage (insofar as

familiarity is computed before activity feeds into the hippocampus).

Because the combined decision emphasizes hippocampal recall, combined recognition degrades along with the hippocampal signal, until the hippocampus is too strongly damaged to produce any signal, at which point combined performance rises to match cortical FC performance. This replicates the Baxter & Murray finding that partial hippocampal lesions can lead to worse recognition performance than complete hippocampal lesions — focusing on the right-hand side of the curve (starting at 75% damage), the size of the hippocampal lesion is negatively correlated with overall FC performance.

As mentioned earlier, the key to understanding the lesion results is that hippocampal damage impairs the hippocampus' ability to carry out pattern separation. We assume that lesioning a layer lowers the total number of units but does not decrease the number of *active* units; as such, percent activity (active units/total units) increases with lesion size — representations become *less sparse* and the average amount of overlap between patterns increases. There is neurobiological support for this assumption: In the brain, the activity of excitatory pyramidal neurons is regulated primarily by inhibitory interneurons (Douglas & Martin, 1990); assuming that both excitatory and inhibitory neurons are damaged by lesions, this loss of inhibition is likely to result in a (proportional) increase in activity for the remaining excitatory neurons. Future work will explore these damage effects in more realistic networks with explicitly-simulated inhibitory interneurons.

As we saw in Simulations 1 and 2 (Figures 7 and 8), low pattern separation switches the hippocampus from a system with (approximate) high-threshold operating characteristics to more of a signal detection process, where both studied items and lures trigger varying degrees of prototype recall, and the two distributions overlap strongly. Thus, as a result of the lesion, the recall signal becomes much more noisy and much less diagnostic.

To back up this claim, we computed the distributions of recall scores triggered by studied items and lures in the intact and lesioned models, and we plotted ROC curves for these models. Figure 37a shows the ROC curves: The intact-model ROC has a positive Y-intercept around .60, whereas the ROC for the 75%-lesioned model has a curvilinear shape and a zero Y-intercept. Figure 37 parts b and c show the underlying recall distributions for the intact and lesioned models, respectively. In the intact model, the lure recall distribution is centered on zero; furthermore, it is possible to set a high recall threshold that is exceeded by (some) studied items but not by lures.

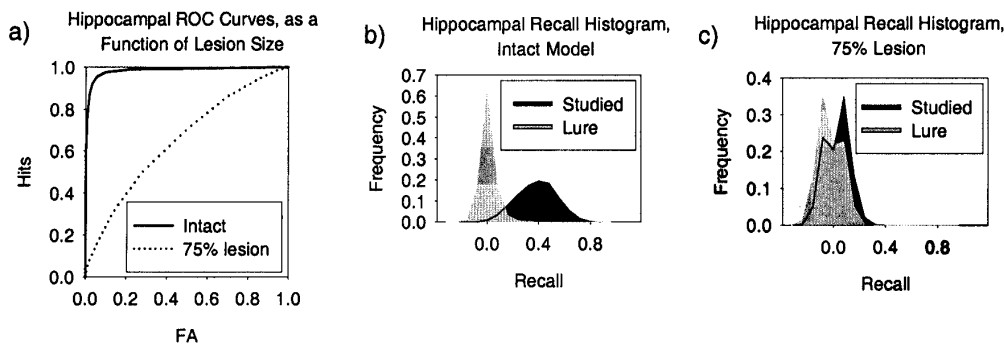


Figure 37: (a) Recall ROC curves for the intact combined model, and after a 75% hippocampal lesion. (b) Histogram showing the distribution of studied and lure recall scores in the intact combined model. (c) Histogram showing the studied and lure recall distributions after a 75% hippocampal lesion. The intact-model ROC has a high Y-intercept (around .60) but the lesioned-model ROC has a zero Y-intercept. The intact-model histogram shows that lures do not trigger recall scores greater than .5, but studied items frequently do. In contrast, the lesioned-model histogram shows a high degree of overlap between studied and lure recall.

#### Effect of Hippocampal Damage on Studied and Lure Recall

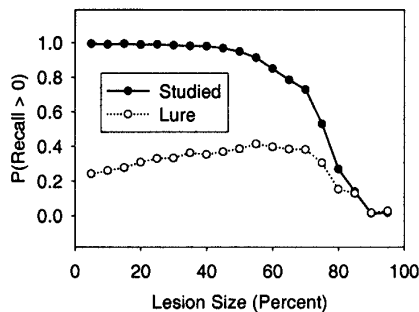


Figure 38: Plot of the probability of studied items and lures triggering above-zero recall, as a function of hippocampal lesion size. For studied items, the probability of above-zero recall declines monotonically as a function of lesion size. For lures, the probability of above-zero recall first increases, then decreases, as a function of lesion size.

In the lesioned model, the studied and lure recall distributions overlap strongly, making it impossible to set a recall threshold that is only exceeded by studied items.

Figure 38 shows that the probability that lures will trigger above-zero recall scores increases steadily with increasing lesion size until it reaches a peak of .41 (for 55% hippocampal damage). Crucially, we hypothesize that lesioned subjects continue to prioritize recall in their recognition decisions even though the presence of recall is not, in fact, highly diagnostic. Overall recognition performance in these subjects suffers because the noisy recall signal drowns out useful information that is present

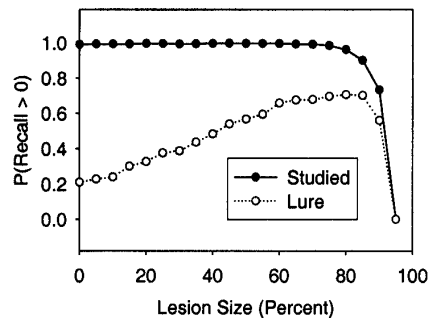
in the familiarity signal. However, as lesion size approaches 60%, the amount of above-zero recall triggered by lures (and studied items) starts to decrease sharply. Thus, larger lesions cause the hippocampus to effectively go silent, and control of the recognition decision reverts to familiarity. This benefits recognition performance insofar as familiarity does a better job of discriminating between studied items and lures than the recall signal generated by the lesioned hippocampus.

#### Region-Specific Hippocampal Lesions

To get at the precise anatomical basis of the effects discussed above, we focally lesioned either DG and CA3 (the hippocampal structures primarily responsible for pattern separation) or CA1 (the hippocampal structure that “decodes” CA3 representations, thereby allowing for recall). Figure 39 shows that lesioning the structures responsible for pattern separation (DG and CA3) leads to a large increase in false recall of lures; false recall continues to increase until lesions are so large that the system stops working altogether (around 85% damage), at which point both studied and lure recall drop precipitously. In contrast, lesioning CA1 leads to a steady decrease in recall associated with studied items and lures.

Thus, the effect of hippocampal damage on recognition can be subdivided into two separate effects: Lesioning the pattern separation apparatus (DG and CA3) adversely affects the diagnosticity of recall, whereas lesioning the translation apparatus (CA1) adversely affects the amount of recall, but not its diagnosticity. When all three structures are lesioned together, the first effect (pattern separation failure) initially predominates, but then the second effect (recall failure) eventually takes over,

a) Effect of DG/CA3 Damage on Studied and Lure Recall



b) Effect of CA1 Damage on Studied and Lure Recall

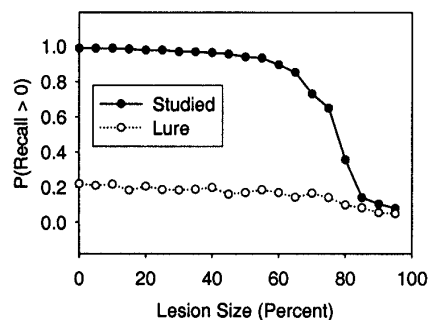


Figure 39: Plot of the probability of studied items and lures triggering above-zero recall, as a function of DG/CA3 and CA1 lesion size. (a) DG/CA3 lesions lead to a sharp increase in (false) recall triggered by lures, until lesion size exceeds 80% (at which point studied-item recall and lure recall both drop off sharply). (b) CA1 lesions lead to a monotonic decrease in the amount of recall triggered by studied items and lures.

removing the harmful effects of noisy recall.

### MTLC Lesions

Turning to the effects of MTLC lesions, we find that lesioning EC.in hurts both cortical and hippocampal recognition performance (Figure 40). Therefore, overall recognition performance decreases steadily as a function of MTLC lesion size. This is exactly the result that was obtained by Baxter & Murray: MTLC lesion size is positively correlated with recognition impairment. The observed deficits result from the fact that overlap between EC.in patterns increases with lesion size — this has a direct adverse effect on cortically-based discrimination; furthermore, because EC.in serves as the input layer for the hippocampus, increased EC.in overlap leads to increased hippocampal overlap, which hurts recall.

Effect of MTLC Damage on FC Recognition

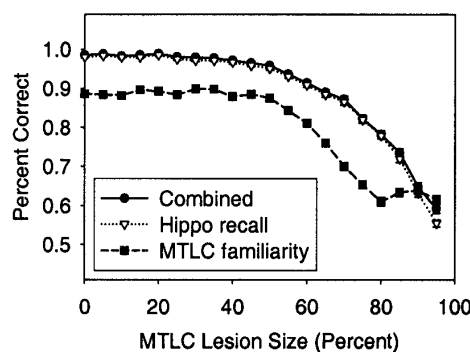


Figure 40: Effect of MTLC (specifically, EC.in) damage on forced-choice (FC) recognition performance. This graph plots FC accuracy based on MTLC familiarity, hippocampal recall, and a combination of the two signals, as a function of EC.in lesion size. All three accuracy scores (recall-alone, familiarity-alone, and combined) decline steadily with increasing lesion size.

### Summary and Implications

In summary, the model was able to reproduce the full pattern of results obtained by Baxter and Murray (2001b) — moving from partial (e.g., 75%) to complete hippocampal lesions improved overall recognition performance, but moving from partial to complete MTLC lesions hurt recognition performance. Most importantly, we were able to provide a principled explanation of why partial hippocampal lesions are so harmful for recognition performance, in terms of pattern separation failure resulting in a noisy recall signal that subjects nevertheless rely upon when making recognition judgments.

We should emphasize that our lesion simulations do not predict complete sparing of recognition performance following complete hippocampal damage. Insofar as recall and familiarity both make independent contributions to recognition, losing recall should always hurt recognition to some degree. But the resulting deficit may be very small and thus hard to detect in single-case studies — the exact size of the deficit will depend on how well the recall signal discriminates between studied items and lures, how strongly correlated recall is with familiarity in this instance, and the exact decision rule that subjects are using.

The finding that partial hippocampal lesions are especially harmful to recognition explains why it is so difficult to find selective sparing of recognition after hippocampal damage (in either humans or monkeys): If the lesion is too small, you end with a partially-lesioned hip-

pocampus that injects noise into the recognition process; if the lesion is too large, then you hit perirhinal cortex (in addition to the hippocampus) and this leads to deficits in familiarity-based recognition. To get selective sparing of recognition, the lesion has to encompass almost all of the hippocampus, and virtually none of the surrounding cortices; the odds of this happening by chance are quite low.

Finally, our model's account of why partial hippocampal lesions impair recognition leads to the following testable prediction. If patients with partial hippocampal lesions show impaired recognition performance because of a noisy recall signal, it should be possible to improve these subjects' recognition performance by using a paradigm that forces them to rely exclusively on familiarity. One way to do this is to use *speeded responding* — recall comes "on-line" later than familiarity because the hippocampus is located downstream of MTL; thus, it should be possible to eliminate the influence of recall on performance by forcing subjects to respond quickly (for examples of how this technique has been used in behavioral studies, see Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; Rotello & Heit, 2000; Hintzman, Caulton, & Levitin, 1998). In summary, we predict that speeded responding should improve recognition in partial-hippocampal-lesion patients, by mitigating recall-related noise. In contrast, speeded responding should hurt recognition in control subjects — recall is a highly diagnostic signal for controls, so removing recall should result in worse performance.

### General Discussion

In the Introduction, we highlighted two important challenges for recognition memory research. The first challenge is to characterize how recall contributes to recognition memory, and how this contribution is different from the contribution of familiarity. The second challenge is to characterize how the hippocampus contributes to recognition memory, and how this contribution is different from the contribution of surrounding medial temporal lobe neocortex (MTLC). Rather than treating these challenges as distinct, we have found considerable synergy in addressing these questions together using a new, biologically-based, dual-process computational model of recognition memory — the *Complementary Learning Systems* (CLS) model. The hippocampal component of this model discriminates between old and new items based on a recall signal, the properties of which depend critically on biologically-motivated pattern separation mechanisms. The neocortical component contributes a non-specific familiarity signal that tracks the *sharpening* of cortical representations that occurs as

a result of Hebbian learning and inhibitory competition. By establishing a clear, mechanistic mapping between the brain structures involved in recognition, and the processes they support, the CLS model makes it possible to bring constraints from neuroscience to bear on the question of how recall and familiarity contribute to recognition. Furthermore, our use of an explicit computational model allows us to describe the respective contributions of hippocampus and neocortex to recognition in a more nuanced fashion.

In what follows, we briefly review the results of our simulations, and then explore the implications of these results for extant theories of recognition memory. We compare our model to other neurally-inspired and abstract memory models. Finally, we discuss some of the limitations of the current model, and our plans for future simulation research.

### Summary of Key Simulation Results

We have identified several fundamental, qualitative differences between the hippocampal recall and MTL familiarity signals. The most important difference is that recall is more diagnostic than familiarity: Assuming low-to-moderate levels of overlap between stimuli, recall behaves in an *approximate high-threshold* fashion — it is possible to set a threshold for matching recall that studied items cross but lures do not. Also, lures sometimes trigger mismatching recall, but studied items do not. Thus, above-threshold levels of matching recall are strong evidence that an item was studied, and mismatching recall is strong evidence that an item was not studied. Approximate high-threshold behavior in the hippocampus is a consequence of the hippocampus' ability to assign distinct representations to stimuli (*pattern separation*) and the thresholded nature of recall, whereby a feature will be recalled only if weights linking that input feature to the output layer were strengthened at study. The hippocampus' tendency to encode feature conjunctions (as opposed to individual features) minimizes the extent to which *blending* of features from different patterns occurs — assuming that average overlap between input patterns is not too high, the hippocampus will only recall features together if these features occurred together study.

We also showed that there are clear boundary conditions on high-threshold responding and low blending in the hippocampus; when hippocampal pattern separation mechanisms fail (because of high input overlap, or as a consequence of damage) the hippocampus degenerates into a state where recall of item-specific features of studied items is poor, and both studied items and lures trigger strong recall of prototype features; these prototype features sometimes are mistakenly recalled in place of item-specific features, resulting in blends. In

this situation, recall resembles a signal detection process more than a high-threshold process. Whereas recall only has signal-detection properties in certain situations, MTLC familiarity always behaves like a signal-detection process — the familiarity distributions associated with studied items and lures are roughly normal and overlap strongly. Because pattern separation is much weaker in cortex than in the hippocampus, the MTLC familiarity signal smoothly tracks the extent to which the test probe matches studied items. Lures that resemble studied items are assigned MTLC representations that overlap strongly with the MTLC representations of studied items; consequently, these lures trigger a strong MTLC familiarity signal.

Recognition tests with related lures (i.e., lures made to resemble specific studied items) reveal a clear dissociation between the two systems: The hippocampus can successfully discriminate between studied items and related lures (because of its superior pattern separation abilities) but cortex is fooled by the strong familiarity signal that these lures generate. However, we also found that use of a *forced-choice* (FC) test format greatly benefits cortical recognition performance with related lures — covariance in the familiarity scores triggered by studied items and corresponding related lures makes the familiarity difference between these items highly reliable. In contrast, hippocampal forced-choice performance (on tests with corresponding related lures) can actually be harmed by covariance: Specifically, covariance between matching recall triggered by studied items, and mismatching recall triggered by related lures, makes it less likely that subjects will be able to reject the lure (due to mismatching recall) in situations where studied recall fails. One way of summarizing the model's predictions is that hippocampally-lesioned patients, who are relying exclusively on MTLC familiarity, should perform poorly relative to controls on standard YN recognition tests with related lures, but they should perform relatively well on FC tests with corresponding related lures, because covariance benefits cortical performance (but not hippocampal performance) on these tests. Holdstock et al. (in press) recently tested a patient with focal hippocampal damage and obtained the predicted pattern of results.

A basic prediction from our model is that the hippocampus should be highly sensitive to conjunctions of input features; cortex should demonstrate some sensitivity, but less than the hippocampus. We demonstrated this using an associative recognition paradigm, where subjects study pairs of stimuli (A-B, C-D) and then have to discriminate studied pairs from recombined pairs (A-D). On a yes-no associative recognition test, hippocampus was clearly better than cortex at rejecting recombined pairs, although cortex nevertheless showed above-floor performance on this task. The model also predicts that

the hippocampus should show less of an advantage on FC associative recognition tests with overlapping lures (i.e., study A-B, C-D; test A-B vs. A-D). We hypothesize that, on these tests, subjects cue recall using the shared pair item (A); this strategy yields suboptimal hippocampal performance, because subjects do not benefit from the pattern separation that occurs when you cue with both parts of a recombined lure, nor do they benefit from recall-to-reject — if the shared cue does not trigger any recall, subjects will not be able to respond correctly based on match (to the studied item) or mismatch (to the lure).

Another key difference between hippocampal and cortical processing is that recognition based on hippocampal recall is more susceptible to interference than recognition based on MTLC familiarity. Interference is typically operationalized in terms of how adding new list items (increasing *list length*) or strengthening memory for some list items (increasing *list strength*) affects memory for other list items. We showed that list strength effects are present for hippocampal recall regardless of input pattern overlap (unless recall is at ceiling) but, for low-to-moderate levels of input overlap, cortex shows a null list strength effect for recognition. The key difference is that, in cortex, interference degrades the model's responding to both studied items and lures, such that (initially) the gap in the model's responding to studied items and lures does not decrease; by contrast, in the hippocampus responding to lures is at floor, so decreased responding to studied items necessarily pushes the studied and lure recall distributions closer together. The studied-lure gap in cortex actually increases slightly with interference, because of *differentiation* — studying an item makes its representation overlap less with the representations of other, interfering items, thus studied items suffer less interference than lures. Importantly, there are limits on this differentiation dynamic; with high levels of interference item strengthening and/or high input overlap, the model's sensitivity to discriminative features of studied items and lures approaches floor and the distributions converge. We presented data from three new list strength experiments (Norman, submitted) that provide support for the model's prediction that list strength affects recall-based recognition but not familiarity-based recognition.

Whereas list strength effects are typically not found in recognition memory experiments (except in the circumstances described by Norman, submitted), list length effects are more reliably obtained. We showed that incorporating biologically-motivated temporal dynamics into our weight change rule allows us to account for this length-strength dissociation using the cortical model: Presenting an item for the first time generates a large (but transient) weight change that produces measurable

interference effects. Repeatedly presenting an item has a small, lasting effect on network weights; however, it does not affect the large, transient component of learning because this component saturates quickly — as such, item repetitions tend not to cause additional interference above and beyond the amount resulting from the item's initial presentation.

Finally, we used our combined cortico-hippocampal model to explore the statistical relationship between recall and familiarity, and the effects of partial hippocampal and cortical lesions. We found that increasing interference lowers the recall-familiarity correlation for studied items — this occurs because interference pushes raw familiarity and recall scores in different directions. Interference is not the only factor that affects the recall-familiarity correlation (e.g., encoding variability and cortical sampling variability boost the recall-familiarity correlation); however, it is the only factor we found that acts to decorrelate the two signals. In the lesion simulations, we found that partial hippocampal lesions can actually lead to worse performance than complete lesions — partial lesions reduce the diagnosticity of the recall signal; if subjects continue to rely heavily on this non-diagnostic signal then performance suffers. These results are consistent with a recent meta-analysis showing a negative correlation between recognition impairment and hippocampal lesion size (Baxter & Murray, 2001b).

### *Comparison with other Theories and Models*

#### *Aggleton & Brown's Recall/Familiarity Theory*

The CLS model is perhaps closest in spirit to the neuropsychological theory of episodic memory set forth by Aggleton and Brown (1999) (A&B). This theory holds that the hippocampus is essential for recall of studied stimuli, following limited exposure to those stimuli, and that MTLC implements a familiarity process that can, in some circumstances, support good discrimination of studied items from nonstudied items (see also Eichenbaum, Otto, & Cohen, 1994, who argue that MTLC is important for "intermediate-term" storage of item information). These are clearly central themes of our CLS model. The main difference between the CLS model and the A&B theory is that the CLS neural network model incorporates specific claims about how the hippocampus implements recall, and how familiarity can be read out from MTLC representations. Moreover, we have implemented these ideas in a working computational model that can be used to test their sufficiency and generate novel predictions. This additional level of mechanistic detail allows the CLS model to provide a more nuanced account of how different manipulations will affect hippocampally-mediated recall and MTLC-mediated familiarity.

For example, A&B assert that the MTLC familiarity process can support intact item recognition performance on its own, whereas memory for associations between items depends on the hippocampal system. However, we showed that the cortex can support a substantial level of associative recognition performance; this explains reports of good associative recognition (using, e.g., unrelated word pair stimuli) in patients with focal hippocampal damage (Vargha-Khadem et al., 1997; Mayes et al., 2001), which are otherwise problematic for the straightforward dichotomy proposed by A&B. Furthermore, the CLS model predicts that cortex will be strongly impaired relative to the hippocampus on certain kinds of item recognition tests (e.g., yes-no tests where lures are similar to studied items); see Holdstock et al. (in press) for evidence consistent with this prediction.

Both our model and the A&B theory predict that item recognition (with unrelated lures) should be spared, relative to recall, following complete hippocampal lesions (as was found, e.g., by Mayes et al., submitted). Thus, it is potentially quite problematic for both models that some studies have found roughly equivalent deficits in recognition and recall following focal hippocampal damage (e.g., Manns & Squire, 1999). We do not claim to fully understand why hippocampal lesions have such variable effects on item recognition; however, it may be possible to account for some of this variability in terms of the idea, set forth by Baxter and Murray (2001b) and explored in simulations here, that partial hippocampal lesions are especially harmful to recognition. This idea is still controversial and needs to be tested directly (e.g., by varying lesion extent and measuring DNMS performance in monkeys).

#### *Yonelinas & Jacoby's Dual-Process Signal Detection Model*

As mentioned earlier, Yonelinas & Jacoby's dual-process signal detection model (Jacoby et al., 1997) is the basis for several widely-used techniques for teasing apart the contributions of recall and familiarity to behavioral recognition performance. This model assumes that familiarity is an equal-variance signal detection process and recall is a high-threshold process (i.e., recall is all-or-none, and only studied items are recalled as being "old"; some variants of the model also make the *dual* high-threshold assumption that lures — but never studied items — are sometimes recalled as being "new"). The Yonelinas & Jacoby model also assumes that recall and familiarity are independent. The CLS model provides some converging support for these assumptions, but it also makes it clear that we should not expect these assumptions to hold under all circumstances. For example, our hippocampal model behaves in a manner that is approximately consistent with dual high-threshold theory

when input overlap is not too high, but it deviates from this behavior in cases of high input overlap or as a result of partial damage. Furthermore, the model's behavior is never strictly consistent with high-threshold theory (e.g., recall is not all-or-none, and lures sometimes trigger above-zero recall scores).

The independence assumption is the most controversial part of the dual-process signal detection model. The CLS model predicts that some degree of positive correlation will be present because the cortical and hippocampal networks are tightly interconnected (i.e., fluctuations in MTLC activity, which are registered as changes in familiarity, affect the extent to which activity propagates through the hippocampus). In our basic-parameter simulations, this "baseline" positive correlation was large (.27) but this may be an artefact of the high degree of sampling variability present in the 240-unit MTLC layer of the combined model. More importantly, the model predicts that situational factors will affect the size of the recall-familiarity correlation: Increasing interference reduces the correlation, but encoding variability boosts the correlation.

In summary: The CLS model allows us to predict, in a principled fashion, when the high-threshold and independence assumptions will hold true. Interestingly, the preconditions for these assumptions being met are, to some extent, in conflict with one another — increasing the average amount of overlap between stimuli boosts interference (which helps foster independence), but too much overlap leads to violations of the high-threshold assumption. Thus, the presence of an intermediate amount of overlap between traces (coupled with low encoding variability) should result in optimal compliance with the assumptions of dual-process signal-detection theory. In this situation, the independence and high-threshold assumptions may not be perfectly valid, but they should hold well enough such that measurement techniques that rely on these assumptions will yield meaningful results.

#### *Bayesian Global Matching Models (e.g., REM)*

We have emphasized how our model spans the gap between neurobiological data and abstract mathematical models, e.g., the REM model developed by Shiffrin & Steyvers, 1997 and its close relative (McClelland & Chappell, 1998). Both of these abstract models carry out a Bayesian computation of the likelihood that an ideal observer should say "old" to an item, based on the extent to which that item matches (and mismatches) stored memory traces. Our cortical familiarity model resembles these abstract models in several respects: Like the abstract models, our cortical model computes a scalar that tracks the "global match" between the test probe and stored memory traces; furthermore, both our cortical model and models like REM attempt to explain the

null LSE for recognition using the differentiation principle introduced by Shiffrin et al. (1990). Thus, our modeling work relies critically on insights that were developed in the context of abstract models like REM.

We can also draw a number of contrasts between the CLS model and abstract Bayesian models. First, our model posits that two processes contribute to recognition whereas these other models attempt to explain recognition data in terms of a single familiarity process. Second, our model incorporates explicit claims about how different brain structures (hippocampus and MTLC) support recognition memory, whereas abstract models do not address how recognition is implemented in the brain. Because of this brain-model mapping, we can constrain our model using neurobiological and neuropsychological data in addition to purely behavioral data. Indeed, there is a strong sense in which the brain-model mapping makes our dual-process approach possible — as discussed in the Introduction, it is not clear how to constrain a dual-process model based on behavioral data alone (because of chicken-and-egg problems). One way of viewing the distinction between abstract and neurobiological recognition memory models is that abstract models characterize recognition memory processes at the *algorithmic* level whereas models like ours focus more on *implementational* details (Marr, 1982). Thus, the two approaches can be viewed as complementary. However, it is not true that the two levels of analysis are completely independent — considering how recognition is implemented in the brain constrains the range of possible algorithms (O'Reilly & Munakata, 2000). Some aspects of current abstract models are highly implausible in light of what we know about brain-style computation, for example the assumption made by REM that memory traces are stored separately, such that no structural interference occurs between memory traces at study.

We have not included head-to-head comparisons between our model and abstract models like REM because our model is, at present, incomplete — as discussed earlier, the model incorporates some sources of variance that we plan to remove and lacks some sources of variance that we plan to add; furthermore, we have not yet settled on a *combined decision rule* (i.e., an algorithm for making recognition decisions based on recall and familiarity). As such, we are not yet in a position to provide precise fits to behavioral recognition memory data. This issue of how subjects utilize memory signals is quite complex, and it lies at the core of the debate over single-process vs. dual-process approaches to memory. No one would deny that recall can contribute to recognition memory; the defining characteristic of single-process approaches is the claim that subjects (for whatever reason) do not utilize recall information that is available. We are committed to the idea that subjects do frequently make



use of recall on recognition tests, but we certainly do not want to claim that subjects always make full use of the recall signal.

In this paper, we have focused on describing qualitative model predictions, and the boundary conditions of these predictions. Working at this level, it is clear that there are some fundamental differences in the predictions of the CLS model vs. models like REM. For example, one key difference between the models is that — in our model — interference occurs at study, when one item re-uses weights that are also used by another item, whereas REM posits that memory traces are stored in a non-interfering fashion, and that interference arises at test, whenever the test item spuriously matches memory traces corresponding to other items. Because studying one item degrades the memory traces of other items, our model predicts — regardless of parameter settings — that the curve relating interference (e.g., list length or list strength) to recognition sensitivity will always asymptotically go to zero with increasing interference. In contrast, in REM it is possible to completely eliminate the deleterious effects of interference items on performance through differentiation; if interference items are presented often enough, they could become so strongly differentiated that the odds of them spuriously matching a test item are effectively zero; whether or not this actually happens depends on model parameters.

#### *Ratcliff's (1990) Neural Network Recognition Model*

Ratcliff (1990) presented a neural network model of recognition memory that is, in some respects, similar to our CLS model. Ratcliff's model is a 3-layer feed-forward network, which learns (using error-driven back-propagation) to reproduce input patterns in the output layer. Like our hippocampal model, Ratcliff's model uses a recall-based dependent measure (i.e., how well does the output pattern match the input pattern), and, like our cortical model, Ratcliff's model uses overlapping distributed representations in its hidden layer. Ratcliff deemed this model to be unsuitable because it shows excessive levels of interference (see also McCloskey & Cohen, 1989).

According to our framework, Ratcliff's model shows excessive interference because it combines two properties (use of a recall dependent measure and overlapping representations) that, in the brain, apply to distinct subsystems. Mixing these properties in one system does not work. In systems that use a recall dependent measure, interference effects on  $d'$  are inevitable because of floor effects on recall of lures — in general, if you did not study a feature, then weights to that feature will not be large enough to support recall; because lure recall is at floor, any decrease in studied recall necessarily pushes the studied and lure recall distributions closer together,

thereby reducing discriminability. Our hippocampal network avoids excessive interference, despite its use of a recall dependent measure, by incorporating automatic pattern-separation mechanisms that reduce overlap between representations in CA3. Our cortical network works in the opposite fashion — it avoids undue interference (despite its use of overlapping representations) because it uses a dependent measure (*act-win*) that is above floor for both studied items and lures; as discussed earlier, interference degrades studied and lure familiarity in tandem, so the studied-lure gap in familiarity is preserved. If we forced our familiarity model to use a recall dependent measure, it would suffer from many of the same problems as Ratcliff's model.

#### *Other Neural Network Models of Hippocampus and Cortex*

The hippocampal component of the CLS model is part of a long tradition of hippocampal modeling (e.g., Marr, 1971; McNaughton & Morris, 1987; Rolls, 1989; Levy, 1989; Touretzky & Redish, 1996; Burgess & O'Keefe, 1996; Wu, Baxter, & Levy, 1996; Treves & Rolls, 1994; Moll & Miikkulainen, 1997; Hasselmo & Wyble, 1997). Although different hippocampal models may differ slightly in the functions they ascribe to particular hippocampal subcomponents, a remarkable consensus has emerged regarding how the hippocampus supports episodic memory (i.e., by assigning minimally overlapping CA3 representations to different episodes, with recurrent connectivity serving to bind together the constituent features of those episodes). In the present modeling work, we build on this shared foundation by applying these biologically-based computational modeling ideas to a rich domain of human memory data (for an application of the same basic model to animal learning data, see O'Reilly & Rudy, 2001).

The Hasselmo and Wyble (1997) model (hereafter, H&W) deserves special consideration because it is the only one of the aforementioned hippocampal models that has been used to simulate patterns of behavioral list-learning data. The architecture of this model is generally similar to the architecture of the CLS hippocampal model, except H&W make a concrete distinction between item and (shared) context information, and posit that item and context information are kept separate throughout the entire hippocampal processing pathway, except in CA3 where recurrent connections allow for item-context associations; furthermore, in the H&W model recognition is based on the extent to which item representations trigger recall of shared contextual information associated with the study list. The H&W model predicts that recognition of studied items should be robust to factors that degrade hippocampal processing because — insofar as all studied items have the same "con-

text vector" — the CA3 representation of shared context information will be very strong and thus easy to activate. However, the fact that the CA3 context representation is easy to activate implies that related lures will very frequently trigger false alarms in the H&W model (in contrast to the CLS model, which predicts low hippocampal false alarms to related lures). The H&W model also predicts a null list strength effect for hippocampally-driven recognition, and a null *main effect* of item strength on hippocampally-driven recognition (in contrast to our model, which predicts that both item strength and list strength effects should be obtained in the hippocampus). Thus, because H&W use a different hippocampal recognition measure, and separate item and context representations, their model generates recognition predictions that are very different from the CLS hippocampal model's predictions. However, we should emphasize that, if H&W used the same recognition measure as our model (match - mismatch), their model and the CLS model would likely make very similar predictions because the two model architectures are so similar.

The neocortical component of the CLS model has much in common with other, recently published neural network models that address the role of cortex in familiarity discrimination (Bogacz, Brown, & Giraud-Carrier, 2001; Sohal & Hasselmo, 2000). These models, like ours, posit that familiarity discrimination in cortex arises from Hebbian learning that tunes a population of units to respond strongly to the stimulus. Both models differ in some ways from ours as well. For example, in the Bogacz et al. (2001) model familiarity is computed by a specialized population of novelty detector units that are not directly involved in representing stimulus properties, whereas our model does not contain specialized novelty detection units — rather, the familiarity signal is computed directly from the activity of units involved in representing the stimulus; at a more detailed level, the Bogacz et al. (2001) model posits that both homosynaptic Hebbian LTD (decrease weights if the sending unit is active but the receiving unit is not) and heterosynaptic Hebbian LTD (decrease weights if the receiving unit is active but the sending unit is not) are important for familiarity discrimination, whereas our model only incorporates heterosynaptic Hebbian LTD. The Sohal and Hasselmo (2000) model, like ours, does not include specialized novelty-detection units, but like the Bogacz model (and unlike ours), it incorporates homosynaptic as well as heterosynaptic Hebbian LTD. For a detailed comparison of the architectural properties of our model vs. the Bogacz et al. (2001) and Sohal and Hasselmo (2000) models, see Bogacz and Brown (submitted).

Perhaps the most salient difference between our modeling work, and the work presented by Bogacz and Sohal & Hasselmo, is that neither Bogacz nor Sohal & Has-

selmo use their models to address detailed patterns of behavioral recognition data; instead, they focus on explaining single-cell recording data in monkeys. While the CLS model can not make detailed predictions about spiking patterns of single neurons, we can account for more general patterns of firing rate changes with familiarity. For example, as discussed earlier, the CLS model predicts that some neurons in perirhinal cortex that initially fire in response to a stimulus will show decreased responding as the stimulus becomes more familiar (e.g., Brown & Xiang, 1998; Li et al., 1993) — these are the neurons that lost the competition to represent the stimulus. In contrast, the neurons that win the competition will not show decreased firing; in our model the activity of these winning neurons actually increases (and we use this increase in *act.win* to index familiarity), but in variants of the model that use more realistic forms of inhibition instead of the k-winners-take-all "shortcut", the activity of winning units does not always increase; in this case, we can read out familiarity in some other way (see the *Alternate Dependent Measures* section below). The model also predicts that neurons that show decreased (vs. asymptotically strong) firing in response to repeated stimulus presentation should be neurons that *initially* had a less strong response to that stimulus. Although there is not space to carry out this analysis here, we think that it would be useful to conduct a detailed comparison of our model's single-cell-firing predictions with the predictions of the other two models, and to determine which model's predictions are most in keeping with the data. We are very open to the possibility that we will have to incorporate additional mechanisms into the model to appropriately fit the single-cell data; furthermore, we realize that adding these mechanisms may alter the model's predictions regarding behavioral recognition performance — one of the strengths of the CLS model is the fact that it provides a conduit whereby low-level neuroscientific results can impact the model's behavioral predictions and vice-versa.

Finally, we should briefly discuss how our model relates to the cortico-hippocampal network model set forth by Gluck and Meyers (e.g., Gluck, Ermita, Oliver, & Meyers, 1997; Gluck & Meyers, 2001). The hippocampal component of the Gluck and Meyers (G&M) model is a three-layer predictive autoencoder network that learns (via error backpropagation) to reproduce the input pattern and predict outcomes on the output layer. G&M have primarily explored hippocampal contributions to conditioning and discrimination learning, and — within this context — they have argued that the primary role of the hippocampus is to pull apart the representations of stimuli that are associated with different outcomes or responses; they call this *predictive differentiation*. The G&M cortical model, by contrast, is not capable of car-

rying out predictive differentiation on its own. Predictive differentiation is a form of pattern separation; however, unlike hippocampal pattern separation in the CLS model, which happens instantly and is automatic, predictive differentiation in the G&M hippocampal model happens over multiple trials, and it only occurs when input patterns are associated with different outcomes/responses. While the G&M model has had considerable success in predicting hippocampal lesion effects in multi-trial discrimination learning tests, we wish to point out that it is not well-suited for modeling episodic memory performance. The G&M hippocampal model is structurally highly similar to the Ratcliff (1990) model discussed earlier (both are three-layer autoencoders that learn via error backpropagation) — like the Ratcliff model, the G&M hippocampal model would suffer from catastrophic interference on episodic memory tests (assuming use of a recall dependent measure) because it does not incorporate an automatic pattern-separation mechanism; also, it would not be able to accommodate any of the other results discussed earlier that rely on automatic pattern-separation in the hippocampus (e.g., superior hippocampal performance on YN related-lure tests). While we can not strongly fault the G&M model for not being able to account for phenomena outside of animal learning, we should point out that the CLS model can account for both animal learning phenomena (O'Reilly & Rudy, 2001) and episodic memory phenomena (as described in this paper) using the same networks.

#### Alternate Dependent Measures

In this research, one important choice we faced was how to apply the cortical and hippocampal networks to recognition, i.e., how do we “read out” signals from these networks that are useful in discriminating between studied and nonstudied items? It is reasonable to ask how much the model's predictions depend on our particular choice of dependent measures. For example, we discussed earlier how the cortical model's robustness to interference (relative to the hippocampus) is due in part to its use of the *act.win* measure as opposed to a “matching recall” measure; are all of the differences between the hippocampus and cortex discussed in this paper due to this difference in how signals are read out from the models?

To address this question, we tested whether the same *act.win* measure used in our cortical model could be used as a recognition signal when applied to area CA1 in the hippocampus. Indeed, we found that this CA1 *act.win* measure yielded respectable  $d'$  scores; crucially, it had the same approximate high-threshold property as the *match – mismatch* hippocampal recall signal (in contrast to the signal-detection property manifested by cortical *act.win*); CA1 *act.win* was more ro-

bust to target-lure similarity manipulations than cortical *act.win*, and it showed a list strength effect on  $d'$  just as our recall signal does. Thus, even when an *act.win* measure is used in both hippocampus and cortex, the key differences noted earlier are still present; as such, the aforementioned differences can not simply be reduced to differences in how we read out signals from the networks.

Another important question has to do with the biological plausibility of the *act.win* measure — it is not immediately clear how some other structure in the brain could isolate the activity of only the winning units (because “losing” units are still active to some small extent, and there are many more losing units than winning units). Therefore, we carried out a fairly exhaustive search through the space of familiarity measures, to see if we could come up with a measure that yields as good or better  $d'$  scores than *act.win*, while also being more biologically plausible. This search yielded one promising measure: the time it takes for activity to spread through the network (*settle.time*). This measure exploits the fact that activity spreads more quickly for familiar vs. unfamiliar patterns. To test this measure, we ran cortical simulations recording *settle.time*, which was operationalized as the number of processing cycles needed for average activity in MTLC to reach a certain threshold (.03). The  $d'$  score computed on this *settle.time* measure was 1.81 ( $SEM = .02$ ), which was comparable to the *act.win*  $d'$  score (2.00,  $SEM = .03$ ). The *settle.time* measure is more biologically plausible than *act.win*, insofar as it only requires some sensitivity to the average activity of a layer, and some ability to assess how much time elapses between stimulus onset and activity reaching a pre-determined threshold.

Finally, we wanted to know if the *settle.time* measure is affected in a manner that is qualitatively similar to *act.win* by key independent variables like list strength. We ran a list strength simulation in the cortical model (list strength was manipulated by doubling the interference-item learning rate from .004 to .008) and computed  $d'$  using *settle.time*. We found that *settle.time*, like *act.win*, does not show a list strength effect on  $d'$  — recognition discrimination was actually better in the strong interference condition (weak interference  $d' = 1.81$ ,  $SEM = .01$ ; strong interference  $d' = 2.05$ ,  $SEM = .01$ ). This is important because it shows that our interference results extend to other dependent measures besides *act.win*. Further research will be necessary to determine if the qualitative properties of *act.win* and *settle.time* are completely identical or if there are manipulations that affect them differently.

### Future Directions

Future research will address limitations of the model that were described earlier. Increases in computer processing speed will make it possible to grow our networks to the point where sampling variability is negligible, and we will replace lost sampling variability by adding encoding variability and variability in pre-experimental presentation frequency to the model. Including pre-experimental variability (by presenting test items in other contexts a variable number of times prior to the start of the experiment) will allow us to address a range of interesting phenomena, including the so-called frequency mirror effect, whereby hits tend to be higher for low-frequency (LF) stimuli than for high-frequency (HF) stimuli, but false alarms tend to be higher for HF stimuli than LF stimuli (see, e.g., Glanzer, Adams, Iverson, & Kim, 1993); recently, several studies have obtained evidence suggesting that recall is responsible for the LF hit-rate advantage and familiarity is responsible for the HF false-alarm-rate advantage (Reder, Nhouyvanisvong, Schunn, Ayers, Angstadt, & Hiraki, 2000; Joordens & Hockley, 2000; Reder et al. also present an abstract dual-process model of this finding).

Furthermore, we plan to directly address the question of how subjects make decisions based on recall and familiarity. Clearly, people are capable of employing a variety of different decision strategies that can differentially weight the different signals that emerge from the cortex and hippocampus. One way to address this issue is to conduct empirical Bayesian analyses to delineate how the optimal way of making recognition decisions in our model varies as a function of situational factors, and then compare the results of these analyses with subjects' actual performance. A specific idea that we plan to explore in detail is that subjects discount recall of prototype information, because prototype recall is much less diagnostic than item-specific recall. The frontal lobes may play an important part in this discounting process — for example, Curran, Schacter, Norman, and Galluccio (1997) studied a frontal-lesioned patient (BG) who false alarmed excessively to nonstudied items that were of the same general type as studied items; one way of explaining this finding is that BG has a selective deficit in discounting prototype recall. Thus, the literature on frontal lesion effects may provide important constraints on how recognition decision-making works, by showing how it breaks down.

Supplementing the model with a more principled theory of how subjects make recognition decisions will make it possible for us to apply the model to a wider range of recognition phenomena, for example situations where recall and familiarity are placed in opposition (e.g., Jacoby, 1991). We could also begin to address

the rich literature on how different manipulations affect recognition ROC curves (e.g., Ratcliff et al., 1992; Yonelinas, 1994).

One other topic for future research involves improving crosstalk between the model and neuroimaging data. In principle, we should be able to predict fMRI activations during episodic recognition tasks by reading out activation from different subregions of the model; to achieve this goal, we will need to build a "back end" onto the model that relates changes in (simulated) neuronal activity to changes in the hemodynamic response that is measured by fMRI. Finally, Curran (2000) has isolated what appear to be distinct ERP correlates of recall and familiarity; as such, we should be able to use the model to predict how these recall and familiarity waveforms will be affected by different manipulations. Our first attempt along these lines was successful; we found that — as predicted by the model — increasing list strength did not affect how well the ERP familiarity correlate discriminates between targets and lures, but list strength adversely affected how well the ERP recall correlate discriminates between targets and lures (Norman, Curran, & Tepe, in preparation).

### Conclusion

We have provided a comprehensive initial treatment of the domain of recognition memory using our biologically-based neural network model of the hippocampus and neocortex. This work extends a similarly comprehensive application of the same basic model to a range of animal learning phenomena (O'Reilly & Rudy, 2001). Thus, we are encouraged by the breadth and depth of data that can be accounted for within our framework. Future work can build upon this solid foundation to address a range of other human and animal memory phenomena.

### Acknowledgements

This work was supported by ONR grant N00014-00-1-0246, NSF grant IBN-9873492, and NIH program project MH47566. KAN was supported by NIH NRSA Fellowship MH12582-01. We thank Tim Curran and David Huber for commenting on the manuscript.

### Appendix A: The Leabra Algorithm

This appendix describes the computational details of the Leabra algorithm that was used in the simulations.

### Pseudocode

The pseudocode for Leabra is given here, showing exactly how the pieces of the algorithm described in more detail in the subsequent sections fit together.

Outer loop: Iterate over events (trials) within an epoch. For each event:

1. Iterate over minus and plus phases of settling for each event.
  - (a) At start of settling, for all units:
    - i. Initialize all state variables (activation,  $v_m$ , etc).
    - ii. Apply external patterns (clamp input in minus, input & output in plus).
  - (b) During each cycle of settling, for all non-clamped units:
    - i. Compute excitatory netinput ( $g_e(t)$  or  $\eta_j$ , eq 7).
    - ii. Compute kWTA inhibition for each layer, based on  $g_i^\Theta$  (eq 10):
      - A. Sort units into two groups based on  $g_i^\Theta$ : top  $k$  and remaining  $k+1$  to  $n$ .
      - B. Set inhib conductance  $g_i$  between  $g_k^\Theta$  and  $g_{k+1}^\Theta$  (eq 9).
    - iii. Compute point-neuron activation combining excitatory input and inhibition (eq 5).
2. Update the weights (based on linear current weight values), for all connections:
  - (a) Compute error-driven weight changes (not used in these simulations).
  - (b) Compute Hebbian weight changes from plus-phase activations (eq 11).
  - (c) Increment the weights according to net weight change, and apply contrast-enhancement (eq 13).

### Point Neuron Activation Function

Leabra uses a *point neuron* activation function that models the electrophysiological properties of real neurons, while simplifying their geometry to a single point. This function is nearly as simple computationally as the standard sigmoidal activation function, but the more biologically-based implementation makes it considerably easier to model inhibitory competition, as described below. Further, using this function enables cognitive models to be more easily related to more physiologically

detailed simulations, thereby facilitating bridge-building between biology and cognition.

The membrane potential  $V_m$  is updated as a function of ionic conductances  $g$  with reversal (driving) potentials  $E$  as follows:

$$\frac{dV_m(t)}{dt} = \tau \sum_c g_c(t) \bar{g}_c (E_c - V_m(t)) \quad (5)$$

with 3 channels ( $c$ ) corresponding to:  $e$  excitatory input;  $l$  leak current; and  $i$  inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component  $g_c(t)$  computed as a function of the dynamic state of the network, and a constant  $\bar{g}_c$  that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential ( $E_e$ ) to 1 and the leak and inhibitory driving potentials ( $E_l$  and  $E_i$ ) of 0:

$$V_m^\infty = \frac{g_e \bar{g}_e}{g_e \bar{g}_e + g_l \bar{g}_l + g_i \bar{g}_i} \quad (6)$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a Bayesian decision making framework (O'Reilly & Munakata, 2000).

The excitatory net input/conductance  $g_e(t)$  or  $\eta_j$  is computed as the proportion of open excitatory channels as a function of sending activations times the weight values:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \quad (7)$$

The inhibitory conductance is computed via the kWTA function described in the next section, and leak is a constant.

Activation communicated to other cells ( $y_j$ ) is a thresholded ( $\Theta$ ) sigmoidal function of the membrane potential with gain parameter  $\gamma$ :

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma[V_m(t) - \Theta]_+}\right)} \quad (8)$$

where  $[x]_+$  is a threshold function that returns 0 if  $x < 0$  and  $x$  if  $x > 0$ . This sharply-thresholded function is convolved with a Gaussian noise kernel ( $\sigma = .005$ ), which reflects the intrinsic processing noise of biological neurons. This produces a less discontinuous deterministic function with a softer threshold that is better suited for graded learning mechanisms (e.g., gradient descent).

### k-Winners-Take-All Inhibition

Leabra uses a kWTA function to achieve sparse distributed representations. Although two different versions

are possible (see O'Reilly & Munakata, 2000 for details), only the simpler, more rigid form was used in the present simulations. A uniform level of inhibitory current for all units in the layer is computed as follows:

$$g_i = g_{k+1}^\Theta + q(g_k^\Theta - g_{k+1}^\Theta) \quad (9)$$

where  $0 < q < 1$  is a parameter for setting the inhibition between the upper bound of  $g_k^\Theta$  and the lower bound of  $g_{k+1}^\Theta$ . These boundary inhibition values are computed as a function of the level of inhibition necessary to keep a unit right at threshold:

$$g_i^\Theta = \frac{g_e^* \bar{g}_e (E_e - \Theta) + g_l \bar{g}_l (E_l - \Theta)}{\Theta - E_i} \quad (10)$$

where  $g_e^*$  is the excitatory net input without the bias weight contribution — this allows the bias weights to override the kWTA constraint.

In the basic version of the kWTA function used here, which is relatively rigid about the kWTA constraint,  $g_k^\Theta$  and  $g_{k+1}^\Theta$  are set to the threshold inhibition value for the  $k$ th and  $k+1$ th most excited units, respectively. Thus, the inhibition is placed exactly to allow  $k$  units to be above threshold, and the remainder below threshold. For this version, the  $q$  parameter is almost always .25, allowing the  $k$ th unit to be sufficiently above the inhibitory threshold.

Activation dynamics similar to those produced by the kWTA function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition (O'Reilly & Munakata, 2000). Thus, although the kWTA function is somewhat biologically implausible in its implementation (e.g., requiring global information about activation states and using sorting mechanisms), it provides a computationally effective approximation to biologically plausible inhibitory dynamics.

### Hebbian Learning

The simplest form of Hebbian learning adjusts the weights in proportion to the product of the sending ( $x_i$ ) and receiving ( $y_j$ ) unit activations:  $\Delta w_{ij} = x_i y_j$ . The weight vector is dominated by the principal eigenvector of the pairwise correlation matrix of the input, but it also grows without bound. Leabra uses essentially the same learning rule used in competitive learning or mixtures-of-Gaussians (Rumelhart & Zipser, 1986; Nowlan, 1990), which can be seen as a variant of the Oja normalization (Oja, 1982):

$$\Delta_{\text{hebb}} w_{ij} = x_i^+ y_j^+ - y_j^+ w_{ij} = y_j^+ (x_i^+ - w_{ij}) \quad (11)$$

Rumelhart and Zipser (1986) and O'Reilly and Munakata (2000) showed that, when activations are interpreted as probabilities, this equation converges on the

conditional probability that the sender is active given that the receiver is active.

To renormalize Hebbian learning for sparse input activations, equation 11 can be re-written as follows:

$$\Delta w_{ij} = \epsilon [y_j x_i (m - w_{ij}) + y_j (1 - x_i) (0 - w_{ij})] \quad (12)$$

where an  $m$  value of 1 gives equation 11, while a larger value can ensure that the weight value between uncorrelated but sparsely active units is around .5. Specifically, we set  $m = \frac{.5}{\alpha_m}$  and  $\alpha_m = .5 - q_m(.5 - \alpha)$ , where  $\alpha$  is the sending layer's expected activation level, and  $q_m$  (called *savg.cor* in the simulator) is the extent to which this sending layer's average activation is fully corrected for ( $q_m = 1$  gives full correction, and  $q_m = 0$  yields no correction).

### Weight Contrast Enhancement

One limitation of the Hebbian learning algorithm is that the weights linearly reflect the strength of the conditional probability. This linearity can limit the network's ability to focus on only the strongest correlations, while ignoring weaker ones. To remedy this limitation, we introduce a contrast enhancement function that magnifies the stronger weights and shrinks the smaller ones in a parametric, continuous fashion. This contrast enhancement is achieved by passing the linear weight values computed by the learning rule through a sigmoidal nonlinearity of the following form:

$$\hat{w}_{ij} = \frac{1}{1 + \left( \theta \frac{w_{ij}}{1 - w_{ij}} \right)^{-\gamma}} \quad (13)$$

where  $\hat{w}_{ij}$  is the contrast-enhanced weight value, and the sigmoidal function is parameterized by an offset  $\theta$  and a gain  $\gamma$  (standard defaults of 1.25 and 6, respectively, used here).

### Appendix B: Basic Parameters

20 items at study: 10 target items (which are tested)  
followed by 10 interference items (which are not tested)  
20% overlap between input patterns (flip 16/24 slots)  
Fixed *high recall threshold*, recall = .40

The following are other basic parameters, most of which are standard default parameter values for Leabra:

Parameter	Value	Parameter	Value
$E_l$	0.15	$\bar{g}_l$	0.235
$E_i$	0.15	$\bar{g}_i$	1.0
$E_e$	1.00	$\bar{g}_e$	1.0
$V_{rest}$	0.15	$\Theta$	0.25
$\tau$	.02	$\gamma$	600
MTLC $\epsilon$	.004	Hippo $\epsilon$	.01
MTLC savg_cor	.4	Hippo savg_cor l	

### Appendix C: Fast Weight Mechanisms

The fast weight values used in the list length simulations are computed by adding in an extra weight value  $f$  to the normal weight value  $w$  (computed just as in the standard Leabra model), using a scaling term  $\lambda$ , to produce an overall weight value  $W$ :

$$W_{ij} = w_{ij} + \lambda f_{ij} \quad (14)$$

The fast weight value is also updated just as the normal weight value is, with two exceptions. First, an offset of .5 is subtracted from the fast weight value so that it is naturally centered around 0 and maintained in the range between -.5 and .5, so that when it is added to the regular weight value it acts like an offset from the normal weight value. Second, the fast weight value decays back to zero as a multiplicative function of its current value (i.e., exponentially):

$$f_{ij}(t+1) = f_{ij}(t) + \epsilon_f [\Delta f_{ij} - d f_{ij}(t)] \quad (15)$$

Where  $\epsilon_f$  is the fast weight learning rate,  $d$  is the decay rate, and  $\Delta f_{ij}$  is the weight change computed just as  $\Delta w_{ij}$ .

The values of these parameters in the simulation were:  $\lambda = .009$ ,  $\epsilon_f = 1.0$ ,  $d = .02$ .

## References

- Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, 22, 425–490.
- Aggleton, J. P., & Shaw, C. (1996). Amnesia and recognition memory: A re-analysis of psychometric data. *Neuropsychologia*, 34, 51–62.
- Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York: Wiley.
- Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., & Lin, L.-H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progress in Brain Research*, 83, 287–300.
- Baum, K. (1997). The list-strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin and Review*, 4, 260–264.
- Baxter, M. G., & Murray, E. A. (2001a). Effects of hippocampal lesions on delayed nonmatching-to-sample in monkeys: A reply to Zola and Squire (2001). *Hippocampus*, 11, 201–203.
- Baxter, M. G., & Murray, E. A. (2001b). Opposite relationship of hippocampal and rhinal cortex damage to delayed nonmatching-to-sample deficits in monkeys. *Hippocampus*, 11, 61–71.
- Beason-Held, L. L., Rosene, D. L., Killiany, R. J., & Moss, M. B. (1999). Hippocampal formation lesions produce memory impairments in the rhesus monkey. *Hippocampus*, 9, 562–574.
- Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361, 31–39.
- Bogacz, R., & Brown, M. W. (submitted). Comparison of computational models of familiarity discrimination in perirhinal cortex.
- Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience*, 10, 5–23.
- Boss, B. D., Peterson, G. M., & Cowan, W. M. (1985). On the numbers of neurons in the dentate gyrus of the rat. *Brain Research*, 338, 144–150.
- Boss, B. D., Turlejski, K., Stanfield, B. B., & Cowan, W. M. (1987). On the numbers of neurons in fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Wistar rats. *Brain Research*, 406, 280–287.
- Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, 2, 51–61.
- Brown, M. W., & Xiang, J. Z. (1998). Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Progress in Neurobiology*, 55, 149–189.
- Burgess, N., & O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus*, 6, 749–762.
- Carpenter, G. A., & Grossberg, S. (1993). Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences*, 16, 131–137.
- Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and Review*, 3, 37–60.
- Clark, S. E., Hori, A., & Callan, D. E. (1993). Forced-choice associative recognition: Implications for global-memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 871–881.
- Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.
- Cohen, N. J., Poldrack, R. A., & Eichenbaum, H. (1997). Memory for items and memory for relations in the procedural/declarative memory framework. In A. R. Mayes, & J. J. Downes (Eds.), *Theories of organic amnesia* (pp. 131–178). Hove, UK: Psychology Press.
- Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory and Cognition*, 28, 923.
- Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 531–547.
- Curran, T., Schacter, D. L., Norman, K., & Galluccio, L. (1997). False recognition after a right frontal lobe infarction: Memory for general and specific information. *Neuropsychologia*, 35, 1035.
- Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108, 452–477.
- Donaldson, W. (1993). Accuracy of d' and a' as estimates of sensitivity. *Bulletin of the Psychonomic Society*, 31, 271–274.
- Douglas, R. J., & Martin, K. A. C. (1990). Neocortex. In G. M. Shepherd (Ed.), *The synaptic organization*



- of the brain (Chap. 12, pp. 389–438). Oxford: Oxford University Press.
- Eichenbaum, H. (2000). Cortical-hippocampal networks for declarative memory. *Nature Reviews Neuroscience*, 1, 41–50.
- Eichenbaum, H., Otto, T., & Cohen, N. J. (1994). Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences*, 17(3), 449–518.
- Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91, 1–67.
- Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, 100, 546–567.
- Gluck, M. A., Ermita, B. R., Oliver, L. M., & Myers, C. E. (1997). Extending models of hippocampal function in animal conditioning to human amnesia. In A. R. Mayes, & J. J. Downes (Eds.), *Theories of organic amnesia* (pp. 179–212). Hove, UK: Psychology Press.
- Gluck, M. A., & Meyers, C. E. (2001). *Gateway to memory: An introduction to neural network modeling of the hippocampus and learning*. Cambridge, MA: MIT Press.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 846–858.
- Hasselmo, M. E. (1995). Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioural Brain Research*, 67, 1–27.
- Hasselmo, M. E., & Wyble, B. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, 89, 1–34.
- Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.
- Hinton, G. E., & Plaut, D. C. (1987). Using fast weights to deblur old memories. *Proceedings of the 9th Annual Conference of the Cognitive Science Society* (pp. 177–186). Hillsdale, NJ: Erlbaum.
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95, 528–551.
- Hintzman, D. L. (2001). Similarity, global matching, and judgments of frequency. *Memory and Cognition*, 29, 547–556.
- Hintzman, D. L., Caulton, D. A., & Levitin, D. J. (1998). Retrieval dynamics in recognition and list discrimination: Further evidence of separate process of familiarity and recall. *Memory and Cognition*, 23, 449–462.
- Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, 33, 1–18.
- Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 667–680.
- Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., & Norman, K. A. (in press). How recall and recognition are affected by focal damage to the human hippocampus. *Hippocampus*.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541.
- Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen, & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 13–47). Mahway, NJ: Lawrence Erlbaum Associates.
- Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1534.
- Kahana, M. J. (submitted). A factorial analysis of the recognition-recall relation in four distributed memory models.
- 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.
- Levy, W. B. (1989). A computational approach to hippocampal function. In R. D. Hawkins, & G. H. Bower (Eds.), *Computational models of learning in simple neural systems* (pp. 243–304). San Diego, CA: Academic Press.
- Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of Neurophysiology*, 69, 1918–1929.

- Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
- Malenka, R. C., & Nicoll, R. A. (1993). NMDA receptor-dependent synaptic plasticity: Multiple forms and mechanisms. *Trends in Neurosciences*, 16, 521-527.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87, 252-271.
- Manns, J. R., & Squire, L. R. (1999). Impaired recognition memory on the Doors and People test after damage limited to the hippocampal region. *Hippocampus*, 9, 495-499.
- Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (London) B*, 262, 23-81.
- Marr, D. (1982). *Vision*. New York: Freeman.
- Mayes, A. R., Holdstock, J. S., Isaac, C. L., Hunkin, N. M., & Roberts, N. (submitted). Relative sparing of item recognition memory in a patient with adult-onset damage limited to the hippocampus.
- Mayes, A. R., Isaac, C. L., Downes, J. J., Holdstock, J. S., Hunkin, N. M., Montaldi, D., MacDonald, C., Cezayirli, E., & Roberts, J. N. (2001). Memory for single items, word pairs, and temporal order in a patient with selective hippocampal lesions. *Cognitive Neuropsychology*, 18, 97-123.
- McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105, 724.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419-457.
- McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation*, vol. 24 (pp. 109-164). San Diego, CA: Academic Press.
- McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, 10(10), 408-415.
- Miller, E. K., Li, L., & Desimone, R. (1991). A neural mechanism for working and recognition memory in inferior temporal cortex. *Science*, 254, 1377-9.
- Mishkin, M., Suzuki, W., Gadian, D. G., & Vargha-Khadem, F. (1997). Hierarchical organization of cognitive memory. *Philosophical Transactions of the Royal Society, London, B*, 352, 1461-1467.
- Mishkin, M., Vargha-Khadem, F., & Gadian, D. G. (1998). Amnesia and the organization of the hippocampal system. *Hippocampus*, 8, 212-216.
- Moll, M., & Miikkulainen, R. (1997). Convergence-zone episodic memory: Analysis and simulations. *Neural Networks*, 10, 1017-1036.
- Murnane, K., & Shiffrin, R. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 17, 855-874.
- Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory and Cognition*, 19, 119-130.
- Murray, E. A., & Mishkin, M. (1998). Object recognition and location memory in monkeys with excitotoxic lesions of the amygdala and hippocampus. *Journal of Neuroscience*, 18, 6568.
- Nicoll, R. A., Kauer, J. A., & Malenka, R. C. (1988). The current excitement in long-term potentiation. *Neuron*, 1, 97-103.
- Norman, K., Curran, T., & Tepe, K. (in preparation). Interference effects on ERP correlates of recall and familiarity.
- Norman, K. A. (submitted). Differential effects of list strength on recollection and familiarity.
- Nowlan, S. J. (1990). Maximum likelihood competitive learning. In D. S. Touretzky (Ed.), *Advances in neural information processing systems*, 2 (pp. 574-582). San Mateo, CA: Morgan Kaufmann.
- Nyberg, L., & Cabeza, R. (2000). Brain imaging of memory. In E. Tulving, & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 501-520). New York: Oxford University Press.
- Ohrt, D. D., & Gronlund, S. D. (1999). List length effect and continuous memory: Confounds and solutions. In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15, 267-273.
- O'Reilly, R. C. (1996). *The Leabra model of neural interactions and learning in the neocortex*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

- O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2(11), 455-462.
- O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, 4(6), 661-682.
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
- O'Reilly, R. C., Norman, K. A., & McClelland, J. L. (1998). A hippocampal model of recognition memory. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems 10* (pp. 73-79). Cambridge, MA: MIT Press.
- O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, 108, 311-345.
- Raaijmakers, J. G., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology*, 43, 205-234.
- Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97, 285-308.
- Ratcliff, R., Clark, S., & Shiffrin, R. (1990). The list strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning Memory and Cognition*, 16, 163-178.
- Ratcliff, R., & McKoon, G. (2000). Memory models. In E. Tulving, & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 571-581). New York: Oxford University Press.
- Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using roc curves. *Psychological Review*, 99, 518-535.
- Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General*, 124, 352-374.
- Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. A. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 294-320.
- Reed, J. M., Hamann, S. B., Stefanacci, L., & Squire, L. R. (1997). When amnesic patients perform well on recognition memory tests. *Behavioral Neuroscience*, 111, 1163-1170.
- Reed, J. M., & Squire, L. R. (1997). Impaired recognition memory in patients with lesions limited to the hippocampal formation. *Behavioral Neuroscience*, 111, 667-675.
- Rempel-Clower, N. L., Zola, S. M., & Amaral, D. G. (1996). Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. *Journal of Neuroscience*, 16, 5233.
- Riches, I. P., Wilson, F. A., & Brown, M. W. (1991). The effects of visual stimulation and memory on neurons of the hippocampal formation and the neighbouring parahippocampal gyrus and inferior temporal cortex of the primate. *Journal of Neuroscience*, 11, 1763-79.
- Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne, & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 240-265). San Diego, CA: Academic Press.
- Rolls, E. T., Baylis, G. C., Hasselmo, M. E., & Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, 76, 153-164.
- Rotello, C. M. (2000). Recall processes in recognition memory. In D. L. Medin (Ed.), *The psychology of learning and motivation*, Vol. 40 (pp. 183-221). San Diego, CA: Academic Press.
- Rotello, C. M., & Heit, E. (2000). Associative recognition: a case of recall-to-reject processing. *Memory and Cognition*, 28, 907.
- Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language*, 43, 67-88.
- Rudy, J. W., & O'Reilly, R. C. (2001). Conjunctive representations, the hippocampus, and contextual fear conditioning. *Cognitive, Affective, and Behavioral Neuroscience*, 1, 66-82.
- Rudy, J. W., & Sutherland, R. W. (1995). Configural association theory and the hippocampal formation: An appraisal and reconfiguration. *Hippocampus*, 5, 375-389.
- Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 5, pp. 151-193). Cambridge, MA: MIT Press.
- Schacter, D. L., Wagner, A. D., & Buckner, R. L. (2000). Memory systems of 1999. In E. Tulving, & F. I. M.

- Craik (Eds.), *The Oxford handbook of memory* (pp. 627–644). New York: Oxford University Press.
- Shiffrin, R., Ratcliff, R., & Clark, S. (1990). The list strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning Memory and Cognition*, 16, 179–195.
- Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 267–287.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin and Review*, 4, 145–166.
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117(1), 34–50.
- Sohal, V. S., & Hasselmo, M. E. (2000). A model for experience-dependent changes in the responses of inferotemporal neurons. *Network: Computation in Neural Systems*, 11, 169.
- Squire, L. R. (1987). *Memory and brain*. Oxford, England: Oxford University Press.
- Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99, 195–231.
- Squire, L. R., Shimamura, A. P., & Amaral, D. G. (1989). Memory and the hippocampus. In J. H. Byrne, & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 208–239). San Diego, CA: Academic Press.
- Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences*, 93, 13515–13522.
- Teyler, T. J., & Discenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, 100, 147–154.
- Touretzky, D. S., & Redish, A. D. (1996). A theory of rodent navigation based on interacting representations of space. *Hippocampus*, 6, 247–270.
- Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4, 374–392.
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26, 1–12.
- Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, 277, 376–380.
- Wu, X., Baxter, R. A., & Levy, W. B. (1996). Context codes and the effect of noisy learning on a simplified hippocampal CA3 model. *Biological Cybernetics*, 74, 159–165.
- Xiang, J. Z., & Brown, M. W. (1998). Differential encoding of novelty, familiarity, and recency in regions of the anterior temporal lobe. *Neuropharmacology*, 37, 657–676.
- Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1341–1354.
- Yonelinas, A. P. (in press). Consciousness, control, and confidence: The three Cs of recognition memory. *Journal of Experimental Psychology: General*.
- Yonelinas, A. P., Dobbins, I., Szymanski, M. D., & Dhaliwal, H. S. and King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition*, 5, 418–441.
- Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 345–355.
- Yonelinas, A. P., & Jacoby, L. L. (1996). Response bias and the process dissociation procedure. *Journal of Experimental Psychology: General*, 125(4), 422–434.
- Zola, S. M., & Squire, L. R. (2001). Relationship between magnitude of damage to the hippocampus and impaired recognition memory in monkeys. *Hippocampus*, 11, 92–98.
- Zola, S. M., Squire, L. R., Teng, E., Sefanacci, L., Bufalo, E., & Clark, R. E. (2000). Impaired recognition memory in monkeys after damage limited to the hippocampal region. *Journal of Neuroscience*, 20, 451–463.
- Zola-Morgan, S., Squire, L. R., & Amaral, D. G. (1986). Human amnesia and the medial temporal region: Enduring memory impairment following a bilateral lesion limited to field ca1 of the hippocampus. *Journal of Neuroscience*, 6, 2950–2967.