Psychological Review
2003, Vol. 110, No. 4, 611–646


0033-295X/03/$12.00 DOI: 10.1037/0033-295X.110.4.611

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

Kenneth A. Norman and Randall C. O'Reilly
University of Colorado at Boulder

The authors present a computational neural-network model of how the hippocampus and medial temporal lobe cortex (MTLC) contribute to recognition memory. The hippocampal component contributes by recalling studied details. The MTLC component cannot support recall, but one can extract a scalar familiarity signal from MTLC that tracks how well a test item matches studied items. The authors present simulations that establish key differences in the operating characteristics of the hippocampal-recall and MTLC-familiarity signals and identify several manipulations (e.g., target–lure similarity, interference) that differentially affect the 2 signals. They also use the model to address the stochastic relationship between recall and familiarity and the effects of partial versus complete hippocampal lesions on recognition.

Memory can be subdivided according to functional categories (e.g., declarative vs. procedural memory; Cohen & Eichenbaum, 1993; Squire, 1992b) and according to neural structures (e.g., hippocampally dependent vs. nonhippocampally dependent forms of memory). Various attempts have been made to align these functional and neural levels of analysis; for example, Squire (1992b) 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; McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Rudy, 2001). In this article, we advance the complementary-learning-systems model by using it to provide a comprehensive treatment of recognition-memory performance.

In this introductory section, we 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. surrounding cortical regions) contributes to recognition. Then, we show how our modeling approach, which is jointly constrained by behavioral and neuroscientific data, can help resolve these controversies.

Kenneth A. Norman and Randall C. O'Reilly, Department of Psychology, University of Colorado at Boulder.

This work was supported by Office of Naval Research Grant N00014-00-1-0246, National Science Foundation Grant IBN-9873492, and National Institutes of Health (NIH) Program Project MH47566. Kenneth A. Norman was supported by NIH National Research Service Award Fellowship MH12582. We thank Rafal Bogacz, Neil Burgess, Tim Curran, David Huber, Michael Hasselmo, Caren Rotello, and Craig Stark for their very insightful comments on a draft of this manuscript.

Correspondence concerning this article should be addressed to Kenneth A. Norman, who is now at the Department of Psychology, Princeton University, Green Hall, Princeton, New Jersey 08544. E-mail: knorman@princeton.edu

## Dual-Process Controversies

Recognition memory refers to the process of identifying stimuli or situations as having been experienced before, for example, when one recognizes a person one knows 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 the person is or where one knows the person from). Nevertheless, recognition can certainly benefit from recall of specific information—if one can recall that a familiar person at the supermarket is in fact one's veterinarian, that reinforces the feeling that one actually does know this person. Theories that posit that recognition is supported by specific recall as well as by nonspecific feelings of familiarity are called *dual-process* theories (see Yonelinas, 2002, for a thorough review of these theories).

Although it is obvious that recall can (in principle) contribute to recognition judgments, the notion that recall routinely contributes to item-recognition performance is quite controversial. Practically all extant math models of recognition consist of a unitary familiarity process that indexes in a holistic fashion the global match between the test probe and all of the items stored in memory (see, e.g., Gillund & Shiffrin, 1984; Hintzman, 1988; Humphreys, Bain, & Pike, 1989). These familiarity-only models can explain a very wide range of recognition findings (for reviews, see Clark & Gronlund, 1996; Raaijmakers & Shiffrin, 1992; Ratcliff & McKoon, 2000)—even findings that, at first glance, appear to require a recall process (see, e.g., McClelland & Chappell, 1998). Furthermore, the relatively small number of findings that cannot be explained using standard global-matching models tend to come from specialized paradigms like Jacoby's process-dissociation procedure (Jacoby, 1991; see Ratcliff, Van Zandt, & McKoon, 1995, for discussion of when global-matching models can and cannot

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.

Another issue that has hindered the acceptance of dual-process models is the difficulty inherent in measuring the separate contributions of recall and familiarity. Several techniques have been devised for quantitatively estimating how recall and familiarity are contributing to recognition performance (receiver operating characteristic [ROC] analysis, independence remember–know, and process dissociation; see Yonelinas, 2001, 2002, for review and discussion), but all of these techniques rely on a core set of controversial assumptions; for example, they all assume that recall and familiarity are stochastically independent. There are reasons to believe that the independence assumption may not always be valid (see, e.g., Curran & Hintzman, 1995). Furthermore, there is no way to test this assumption using behavioral data alone because of chicken-and-egg problems (i.e., 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 measurement techniques that assume (among other things) independence—these data could potentially be used to constrain dual-process models. However, on the theoretical side, modelers are not making use of these data because of reasonable concerns about the validity of the assumptions used to collect them and because single-process models have been quite successful at explaining recognition data (so why bother with more complex dual-process models?). To resolve this impasse, one needs some source of evidence that one can use to specify the properties of recall and familiarity other than the aforementioned measurement techniques.

## Cognitive-Neuroscience Approaches to Recognition Memory

Just as controversies exist in the math-modeling literature regarding the contribution of recall to recognition memory, parallel controversies exist in the cognitive-neuroscience literature regarding the contribution of the hippocampus to recognition memory.

Researchers have long known that the medial temporal region of the brain is important for recognition memory. Patients with medial temporal lobe lesions encompassing both the hippocampus and surrounding cortical regions (perirhinal, entorhinal, and parahippocampal cortices, which we refer to jointly as medial temporal lobe cortex [MTLC]) typically show impaired recall and recognition but intact performance on other memory tests (e.g., perceptual priming, skill learning; see Squire, 1992a, for a review).

The finding of impaired recall and recognition in medial temporal amnesics is the basis for several influential taxonomies of memory. Most prominently, Squire (1987, 1992b), Eichenbaum and Cohen (Cohen & Eichenbaum, 1993; Cohen, Poldrack, & Eichenbaum, 1997; Eichenbaum, 2000), 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). Researchers have argued that the medial temporal region is important for declarative memory because it is located at the 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 (see, e.g., Mishkin, Suzuki, Gadian, & Vargha-Khadem, 1997; Mishkin, Vargha-Khadem, & Gadian, 1998). See Figure 1 for a schematic diagram of how hippocampus, MTLC, and neocortex are connected.

Although the basic declarative-memory framework is widely accepted, attempts to tease apart the contributions of different medial temporal structures have been more controversial. There is widespread agreement that the hippocampus is critical for recall—focal hippocampal lesions lead to severely impaired recall performance. However, the data are much less clear regarding effects of focal hippocampal damage on recognition. Some studies have found roughly equal impairments in recall and recognition (see, e.g., Manns & Squire, 1999; Reed, Hamann, Stefanacci, & Squire, 1997; Reed & Squire, 1997; Rempel-Clower, Zola, & Amaral, 1996; Zola-Morgan, Morgan, Squire, & Amaral, 1986), whereas other studies have found relatively spared recognition after focal hippocampal lesions (see, e.g., Holdstock et al., 2002; Mayes, Holdstock, Isaac, Hunkin, & Roberts, 2002; Vargha-Khadem et al., 1997). 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 (see, e.g., Murray & Mishkin, 1998), whereas others have found impaired recognition (see, e.g., Beason-Held, Rosene, Killiany, & Moss, 1999; Zola et al., 2000). 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 (see, e.g., Aggleton & Shaw, 1996).

Aggleton and Brown (1999) have tried to frame the difference between hippocampal and cortical contributions in terms of dual-

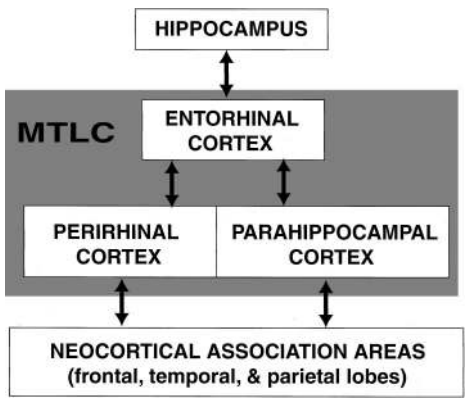


*Figure 1.* Schematic box diagram of neocortex, medial temporal lobe cortex (MTLC), and the hippocampus. MTLC serves as the interface between neocortex and the hippocampus. MTLC is located at the very top of the cortical processing hierarchy—it receives highly processed outputs of domain-specific cortical 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. Adapted from "The Medial Temporal Memory System," by L. R. Squire and S. Zola-Morgan, 1991, *Science, 253,* p. 1380. Copyright 1991 by the American Association for the Advancement of Science (http://www.sciencemag.org). Adapted with permission.

process models of recognition. According to this view, (a) the hippocampus supports recall, and (b) 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 the considerable variability that exists across patients and tests in how hippocampal damage affects recognition. Just as some hippocampal patients have shown more of a recognition deficit than others, some studies have found (within individual patients) greater impairment on some recognition tests than others (see, e.g., Holdstock et al., 2002). In the absence of further specification of the hippocampal contribution (and how it differs from the contribution of MTLC), it is not possible to proactively determine whether recognition will be impaired or spared in a particular patient and/or test.

Aggleton and Brown (1999) attempted to flesh out their theory by arguing that MTLC familiarity can support recognition of individual items but that memory for new associations between items depends on hippocampal recall (Eichenbaum, Otto, & Cohen, 1994, and Sutherland & Rudy, 1989, made similar claims). This view implies that item recognition should be intact but the ability to form new associations should be impaired after focal hippocampal damage. However, Andrew Mayes and colleagues have found that hippocampally lesioned patient Y.R., who has shown intact performance on some item-recognition tests (see, e.g., Mayes et al., 2002), showed impaired performance on other item-recognition tests (see, e.g., Holdstock et al., 2002) and spared performance on some tests that require participants to associate previously unrelated stimuli (e.g., the words *window* and *reason*; Mayes et al., 2001).

In summary, it is becoming increasingly evident that the effects of hippocampal damage are complex. There appears to be some functional specialization in the medial temporal lobe, but the simple dichotomies that have been proposed to explain this specialization either are too vague (recall vs. familiarity) or are inconsistent with recently acquired data (item memory vs. memory for new associations).

## 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 cross talk: 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 to adequately characterize differences in how MTLC contributes to recognition as compared with hippocampus.

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.

Our model falls clearly in the dual-process tradition insofar as we posit that the hippocampus and MTLC contribute signals with distinct properties to recognition memory. The key, differentiating property is that—in our model—differences in the two signals are grounded in architectural differences between the hippocampus and MTLC; because most of these architectural differences fall along a continuum, it follows that differences in the two signals are more nuanced than the dichotomies (item vs. associative) discussed above.

## Precis of Modeling Work

In this section, we summarize the major claims of the article, with pointers to locations in the main text, below, where these issues are discussed in greater detail.

### *Complementary Learning Systems*

The hippocampus is specialized for rapidly memorizing specific events, and the neocortex is specialized for slowly learning about the statistical regularities of the environment. These are the central claims of the complementary-learning-systems (CLS) framework (McClelland et al., 1995; O'Reilly & Rudy, 2001). According to this framework, the two goals of memorizing specific events and learning about statistical regularities are in direct conflict when implemented in neural networks; thus, to avoid making a trade-off, human beings have evolved specialized neural systems for performing these tasks (see Marr, 1971; O'Keefe & Nadel, 1978; Sherry & Schacter, 1987, for similar ideas and Carpenter & Grossberg, 1993, for a contrasting perspective).

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

### *A Dual-Process Model of Recognition*

We have developed models of the hippocampus and neocortex that incorporate key aspects of the biology of these structures and instantiate the CLS principles outlined above (McClelland et al., 1995; O'Reilly & McClelland, 1994; O'Reilly & Munakata, 2000; O'Reilly, Norman, & McClelland, 1998; O'Reilly & Rudy 2001). (see Modeling Framework, below).

The cortical model supports familiarity judgments based on the sharpness of representations in MTLC. Competitive self-organization (arising from Hebbian learning and inhibitory competition) causes stimulus representations to become sharper over repeated exposures (i.e., activity is concentrated in smaller number of units; see The Cortical Model, below). However, the cortical model cannot support recall of details from specific events owing to its relatively low learning rate and its inability to sufficiently differentiate the representations of different events. Familiarity is measured by the activity of the top $k$ most active units, though

other measures are possible, and we have explored some of them, with similar results (for discussion of alternate measures, see *Future Directions,* below).

The hippocampal model supports recall of previously encountered stimuli. When stimuli are presented at study, the hippocampal model assigns relatively nonoverlapping (pattern-separated) representations to these items in region CA3. Active units in CA3 are linked to one another and to a copy of the input pattern (via region CA1). At test, presentation of a partial version of a studied pattern leads to reconstruction (*pattern completion*) of the original CA3 representation and, through this, to reconstruction of the entire studied pattern on the output layer (see The Hippocampal Model, below). The hippocampal model is applied to recognition by computing the degree of match between retrieved information and the recall cue, minus the amount of mismatch; recall of matching information is evidence that the cue was studied, and recall of information that mismatches the retrieval cue is evidence that the cue was not studied.

## Part 1: Basic Network Properties

The signals generated by the hippocampal and cortical models have distinct operating characteristics. This difference is largely due to differences in the two networks' ability to carry out pattern separation (see *Simulation 1: Pattern Separation* and *Simulation 2: Nature of the Underlying Distributions*, below).

MTLC familiarity functions as a standard signal-detection process. In the cortical model, the familiarity distributions for studied items and lures are Gaussian and overlap extensively. The distributions overlap because the representations of studied items and lures overlap in MTLC. Some lures, by chance, have representations that overlap very strongly with the representations of studied items in MTLC, and—as a result—these lures trigger a strong familiarity signal.

In contrast, the hippocampal recall signal is more diagnostic: In the hippocampal model, studied items sometimes trigger strong matching recall, but most lures do not trigger any recall because of the hippocampus's tendency to assign distinct representations to stimuli (regardless of similarity). As such, high levels of matching recall strongly indicate that an item was studied. However, there are boundary conditions on the diagnosticity of the hippocampal recall signal. When the average amount of overlap between stimuli is high, hippocampal pattern-separation mechanisms break down, resulting in strong recall of shared, prototypical features (even in response to lures) and in poor recall of features that are unique to particular studied items.

The difference in operating characteristics between the MTLC familiarity and hippocampal recall signals is most evident on yes–no (YN) *related-lure* recognition tests where lures are similar to studied items but studied items are dissimilar to one another. The hippocampal model strongly outperforms the cortical model on these tests (see *Simulation 3: YN Related-Lure Simulations*, below). As target–lure similarity increases, lure familiarity increases steadily, but (up to a point) hippocampal pattern separation works to keep lure recall at floor. Very similar lures trigger recall, but when this happens the lure can often be rejected due to mismatch between retrieved information and the recall cue. For example, if the model studies *rats* and is tested with *rat*, it might recall having studied *rats* (and reject *rat* on this basis). On related-lure recognition tests, the presence of any mismatching recall is highly diagnostic of the item being a lure. As such, we apply a *recall-to-reject* strategy to hippocampal recall on such tests whereby items are given a confident *new* response if they trigger any mismatching recall.

Finally, in the *Sources of Variability* section, below, we discuss different kinds of variability (e.g., variability due to random initial weight settings, encoding variability) and their implications for recognition performance.

## Part 2: Applications to Behavioral Phenomena

### Interactions Between Lure Relatedness and Test Format

The hippocampal model's advantage for related lures (observed with YN tests, as discussed above) is mitigated by giving the models a forced choice between studied items and corresponding related lures at test (i.e., test A vs. A′, B vs. B′, where A′ and B′ are lures related to A and B, respectively; see *Simulation 4: Lure-Relatedness and Test-Format Interactions*, below). Cortical performance benefits from the high degree of covariance in the familiarity scores triggered by studied items and corresponding related lures, which makes small familiarity differences highly reliable. In contrast, use of this test format may actually harm hippocampal performance relative to other test formats. On forced-choice (FC) tests with noncorresponding lures (test A vs. B′, B vs. A′), the hippocampal model has two independent chances to respond correctly: On trials where it fails to recall the studied item (A), it can still respond correctly if the lure (B′) triggers mismatching recall (and is rejected on this basis). Use of corresponding lures deprives the model of this second chance insofar as recall triggered by studied items and corresponding lures is highly correlated—if A does not trigger recall, A′ will not trigger recall-to-reject.

The predicted interaction between target–lure similarity and test format was obtained in experiments comparing a focal hippocampal amnesic with control participants. The model predicts 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, and they should perform well relative to controls on both YN and FC tests that use unrelated lures (insofar as both networks discriminate well between studied items and unrelated lures). Holdstock et al. (2002) and Mayes et al. (2002) found exactly this pattern of results in hippocampally lesioned patient Y.R.

### Associative Recognition

Associative recognition tests (i.e., study A–B, C–D; test with studied pairs and recombined pairs like A–D) can be viewed as a special case of the related-lure paradigm discussed above (see *Simulation 5: Associative Recognition and Sensitivity to Conjunctions*, below). The hippocampal model outperforms the cortical model on YN associative-recognition tests. As in the related-lure simulations, the hippocampal advantage is due to hippocampal pattern separation, and the hippocampus's ability to carry out recall-to-reject. However, even though the cortical model is worse at associative recognition than the hippocampus, the cortical model still performs well above chance. This finding shows that our

cortical model has some (albeit reduced) sensitivity to whether items occurred together at study, unlike other models (e.g., Rudy & Sutherland, 1989) that posit that cortex supports memory for individual features but not novel feature conjunctions. We also found that giving the models a forced choice between studied pairs and overlapping re-paired lures (test A–B vs. A–D) mitigates the hippocampal advantage for associative recognition. This mirrors the finding that FC testing with corresponding related lures mitigates the hippocampal advantage for related lures. Finally, we present data from previously published studies that support the model's predictions regarding how focal hippocampal damage should affect associative-recognition performance, as a function of test format.

### *Interference Effects*

Hippocampally and cortically driven recognition can be differentially affected by interference manipulations: Increasing *list strength* impairs discrimination of studied items and lures in the hippocampal model but does not impair discrimination based on MTLC familiarity. The list-strength paradigm measures how repeated study of a set of interference items affects participants' ability to discriminate between nonstrengthened (but studied) target items and lures (see *Simulation 6: Interference and List Strength*, below).

In both models, the overall effect of interference is to decrease weights to discriminative features of studied items and lures and to increase weights to prototypical features (which are shared by a high proportion of items in the item set). The hippocampal model predicts a list-strength effect (LSE) because increasing list strength reduces recall of discriminative features of studied items, and lure recall is already at floor. Increasing list strength therefore has the effect of pushing the studied and lure recall distributions together (reducing discriminability). The cortical model predicts a null LSE because the familiarity signal triggered by lures has room to decrease as a function of interference. Increasing list strength reduces responding to (the discriminative features of) both studied items and lures, but the average difference in studied and lure familiarity does not decrease, so discriminability does not suffer.

In the cortical model, lure familiarity initially decreases more than studied-item familiarity as a function of list strength, so the studied–lure gap in familiarity actually widens slightly with increasing interference. The widening of the studied–lure gap can be explained in terms of *differentiation*: Studying an item makes its representation overlap less with the representations of other, interfering items (Shiffrin, Ratcliff, & Clark, 1990); therefore, studied items suffer less interference than lures. However, according to the model, there are limits on this dynamic. With high levels of interference item strengthening and/or high input overlap, the cortical model's sensitivity to discriminative features of studied items and lures approaches floor, and the studied and lure familiarity distributions start to converge (resulting in decreased discriminability).

Data from two recent list-strength experiments provide support for the model's prediction that list strength should affect recall-based discrimination but not familiarity-based discrimination (Norman, 2002). We also discuss ways of modifying the learning rule to accommodate the finding that increasing *list length* (i.e., adding new items to the study list) hurts recognition sensitivity more than increasing list strength (Murnane & Shiffrin, 1991a; see *List-Length Effects*, below).

### *The Combined Model and Independence*

The extent to which the neocortical and hippocampal recognition signals are correlated varies in different conditions. These issues are explored using a more realistic combined model where the cortical network responsible for computing familiarity serves as the input to the hippocampal network (see *Simulation 7: The Combined Model and the Independence Assumption*, below).

In the combined model, variability in how well items are learned at study (encoding variability) bolsters the correlation between recall and familiarity signals, as was postulated by Curran and Hintzman (1995) and others. In contrast, increasing interference reduces the recall–familiarity correlation for studied items. This occurs because interference pushes raw recall and familiarity scores in different directions (increasing interference reduces recall, but asymptotically it boosts familiarity by boosting the model's responding to shared, prototypical-item features). Taken together, these results show that recall and familiarity can be independent when there is enough interference to counteract the effects of encoding variability.

### *Effects of Partial Lesions*

Partial hippocampal lesions can lead to worse overall recognition performance than complete lesions (see *Simulation 8: Lesion Effects in the Combined Model*, below). In the hippocampal model, partial lesions cause pattern-separation failure, which sharply reduces the diagnosticity of the recall signal. Recognition performance suffers because the noisy recall signal drowns out useful information that is present in the familiarity signal. Moving from a partial hippocampal lesion to a complete lesion improves performance by removing this source of noise. In contrast, increasing MTLC lesion size in the model leads to a monotonic decrease in performance. This occurs because MTLC lesions directly impair familiarity-based discrimination and indirectly impair recall-based discrimination (because MTLC serves as the input to the hippocampus). These results are consistent with a recent meta-analysis of the lesion data showing a negative correlation between recognition impairment and hippocampal lesion size and a positive correlation between recognition impairment and MTLC lesion size (Baxter & Murray, 2001b).

## A Note on Terminology

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 signal because it adheres to the definition of familiarity set forth by Hintzman (1988), Gillund and Shiffrin (1984), and others, that is, it is a scalar that tracks the global match or similarity of the test probe to studied items.

However, we realize that the terms *recall* and *familiarity* come with a substantial amount of theoretical baggage. Over the years, researchers have made a very large number of claims regarding

properties of recall and familiarity; for example, Yonelinas has argued that recall is a high-threshold process (see, e.g., Yonelinas, 2001), and Mandler (1980) and others (e.g., Aggleton & Brown, 1999) have argued that familiarity reflects memory for individual items apart from their associations with other items and contexts. By linking the hippocampal contribution with recall and the MTLC contribution with familiarity, we do not mean to say that all of the various (and sometimes contradictory) properties that have been ascribed to recall and familiarity over the years apply to the hippocampal and MTLC contributions, respectively. Just the opposite: We hope to redefine the properties of recall and familiarity using neurobiological data on the properties of the hippocampus and MTLC. In this article, we systematically delineate how the CLS model's claims about hippocampal recall and MTLC familiarity deviate from claims made by existing dual-process theories.

## Modeling Framework

Both the hippocampal and neocortical networks utilize the Hebbian component of O'Reilly's Leabra algorithm (O'Reilly, 1996, 1998; O'Reilly & Munakata, 2000; the full version of Leabra also incorporates error-driven learning, but error-driven learning was turned off in the simulations reported here). The algorithm we used incorporates several widely accepted characteristics of neural computation, including Hebbian long-term potentiation/long-term depression (LTP/LTD) and inhibitory competition between neurons, that were first brought together by Grossberg (1976). (For more information on these mechanisms, see also Kanerva, 1988; Kohonen, 1977; Minai & Levy, 1994; Oja, 1982; Rumelhart & Zipser, 1986.) In our model, LTP is implemented by strengthening the connection (weight) between two units when both the sending and receiving units are active together; 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 using a *k*-winners-take-all (*k*WTA) algorithm, which sets the amount of inhibition for a given layer such that at most *k* units are strongly active. Although the *k*WTA rule sets a firm limit on the number of units that show strong (>.25) activity, there is still considerable flexibility in the overall distribution of activity across units in a layer. This is important for our discussion of sharpening in the cortical model, below. Key aspects of the algorithm are summarized in Appendix A.

## The Cortical Model

The cortical network is composed of two layers, input (which corresponds to cortical areas that feed into MTLC) and hidden (corresponding to MTLC; see Figure 2). 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 twenty-four 10-unit slots, with 1 unit out of 10 active in each slot. A useful way to think of slots is that different slots correspond to different feature dimensions (e.g., color or shape) and different units within a slot correspond to different, mutually exclusive features along that dimension (e.g., shapes: circle, square, triangle). The hidden (MTLC) layer consists of 1,920 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.

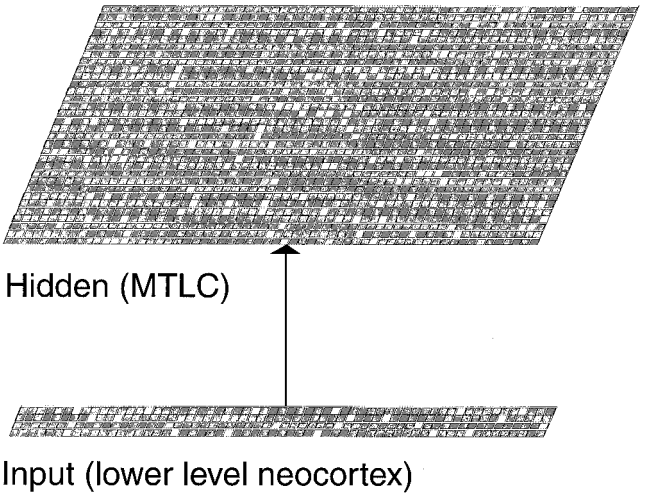


*Figure 2.* Diagram of the cortical network. The cortical network consists of two layers, an input layer (corresponding to lower cortical regions that feed into medial temporal lobe cortex [MTLC]) and a hidden layer (corresponding to MTLC). Units in the hidden layer compete to encode (via Hebbian learning) regularities that are present in the input layer.

Input patterns were constructed from prototypes by randomly selecting a feature value (possibly identical to the prototype feature value) for a random subset of slots. The number of slots that are flipped (i.e., given a random 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 on average (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 nonprototypical, 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 one studies pictures of people from Norway, one sees that most people there 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 are viewed on a particular monitor in a particular room on a particular day). Some simulations involved more complex stimulus construction, as described where applicable.

To apply the cortical model to recognition, we exploited the fact that—as items are presented repeatedly—their representations in the MTLC layer become sharper (see Figure 3). 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

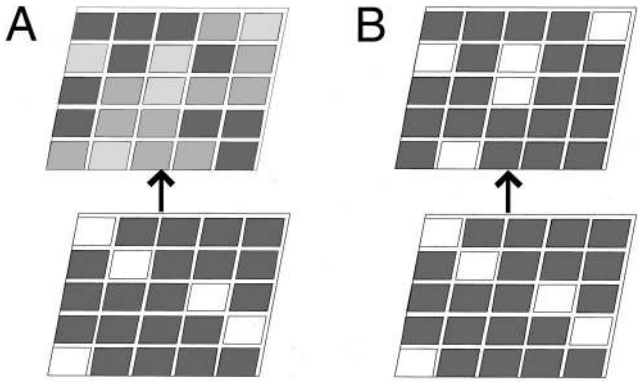


*Figure 3.* Illustration of the sharpening of hidden (medial temporal lobe cortex [MTLC]) layer activation patterns in a miniature version of our cortical model. A: The network prior to sharpening; MTLC activations (more active = lighter color) are relatively undifferentiated. B: The network after Hebbian learning and inhibitory competition produce sharpening; a subset of the units are strongly active, with the remainder inhibited. In this example, we would read out familiarity by measuring the average activity of the $k = 5$ most active units.

chance respond more strongly to the stimulus than other units; these units get tuned by Hebbian learning to respond even more strongly to the item the next time it is presented, and these strongly active units start to inhibit units that are less strongly active (for additional discussion of the idea that familiarization causes some units to drop out of the stimulus representation, see Li, Miller, & Desimone, 1993). We should note that sharpening is not a novel property of our model—rather, it is a general property of competitive-learning networks with graded unit-activation values in which there is some kind of contrast enhancement within a layer (see, e.g., Grossberg, 1986, Section 23; Grossberg & Stone, 1986, Section 16).

The sharpening dynamic in our model is consistent with neural data on the effects of repeated presentation of stimuli in cortex. Electrophysiological studies have shown that some neurons that initially respond to a stimulus exhibit a lasting decrease in firing whereas other neurons that initially respond to the stimulus do not exhibit decreased firing (see, e.g., Brown & Xiang, 1998; Li et al., 1993; Miller, Li, & Desimone, 1991; Riches, Wilson, & Brown, 1991; Rolls, Baylis, Hasselmo, & Nalwa, 1989; Xiang & Brown, 1998). According to our model, the 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 measured the average activity of the MTLC units that won the competition to represent the stimulus. That is, we took the average activation of the top $k$ (192 or 10% of the MTLC) units computed according to the $k$WTA 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 is therefore a function of the exact balance between these increases and decreases (for an example of another model that bases recognition decisions on an activity readout from a neural network doing competitive learning, see Grossberg & Stone, 1986).

Although we used *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 it is far from the only way of operationalizing familiarity, and it is unclear how other brain structures might read out *act_win* from MTLC. We briefly describe another, more neurally plausible familiarity measure (the time it takes for activation to spread through the network) in the General Discussion section.

Finally, we should point out that the idea (espoused above) that the same network is involved in feature extraction and familiarity discrimination is controversial; in particular, Malcolm Brown, Rafal Bogacz, and their colleagues (see, e.g., Bogacz & Brown, 2003; Brown & Xiang, 1998) have argued that specialized populations of neurons in MTLC are involved in feature extraction versus familiarity discrimination. At this point, it suffices to say that our focus in this article is on the familiarity-discrimination capabilities of the cortical network rather than its ability to extract features. We address Brown and Bogacz's claims in more detail in the General Discussion.

## The Hippocampal Model

We have developed a standard model of the hippocampus (O'Reilly & Munakata, 2000; O'Reilly et al., 1998; O'Reilly & Rudy, 2001; Rudy & O'Reilly, 2001) that implements widely accepted ideas of hippocampal function (Hasselmo, 1995; Hebb, 1949; Marr, 1971; McClelland et al., 1995; McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Rolls, 1989). Our goal in this section is to describe the model in just enough detail to motivate the model's predictions about recognition memory. Additional details regarding the architecture of the hippocampal model (e.g., the percentage activity in each layer of the model) are provided in Appendix B.

In the brain, entorhinal cortex (EC) is the interface between hippocampus and neocortex; superficial layers of EC send input to the hippocampus, and deep layers of EC 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. Like the input layer of the cortical model, both EC_in and EC_out have a slotted structure (twenty-four 10-unit slots, with 1 unit per slot active).

Figure 4 shows the structure of the model. The job of the hippocampal model, stated succinctly, is to store patterns of EC_in activity in a manner that supports subsequent recall of these patterns on EC_out. The hippocampal model achieves this goal in the following stages: Input patterns are presented to the model by clamping those patterns onto the input layer, which serves to impose the pattern on EC_in via fixed, one-to-one connections. From EC_in, activation spreads both directly and indirectly (via the dentate gyrus [DG]) to region CA3. The resulting pattern of activity in CA3 is stored by Hebbian weight changes in the feedforward pathway and by strengthening recurrent connections

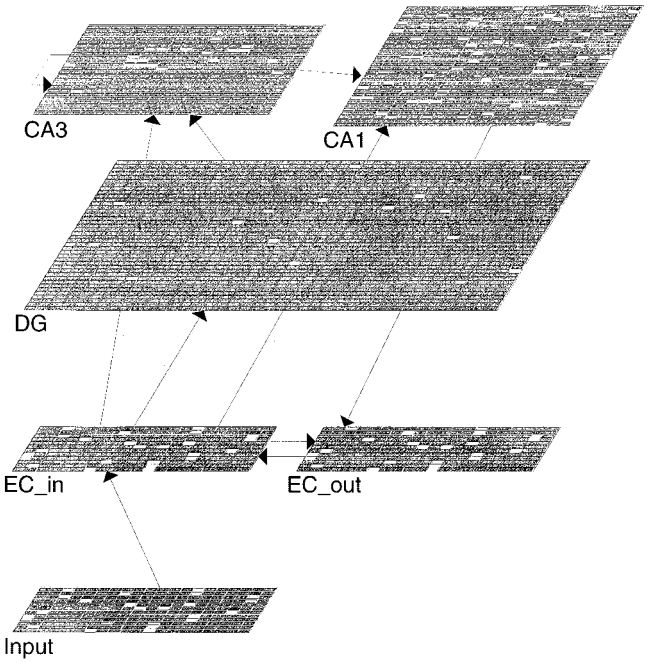


*Figure 4.* Diagram of the hippocampal network. The hippocampal network links input patterns in entorhinal cortex (EC) to relatively nonoverlapping (pattern-separated) sets of units in region CA3; recurrent connections in CA3 bind together all of the units involved in representing a particular EC pattern; 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 on the basis of partial cues. The dentate gyrus (DG) serves to facilitate pattern separation in region CA3.

in CA3 between active units; these weight changes serve to bind the disparate elements of the input pattern by linking them to a shared episodic representation in CA3.

An important property of DG and CA3 is that representations in these structures are very sparse—relatively few units are active for a given stimulus. The hippocampus's use of sparse representations gives rise to pattern separation. If only a few units are active per input pattern, then overlap between the hippocampal representations of different items tends to be minimal (Marr, 1971; see O'Reilly & McClelland, 1994, for a mathematical analysis of how sparseness results in pattern separation and the role of the DG in facilitating pattern separation).

Next, to complete the loop, the CA3 representation needs to be linked back to the original input pattern. This is accomplished by linking the CA3 representation to active units in region CA1. Like CA3, region CA1 contains a re-representation of the input pattern. However, unlike the CA3 representation, the CA1 representation is invertible—if an item's representation is activated in CA1, well-established connections between CA1 and EC_out allow activity to spread back to the item's representation in EC_out. Thus, CA1 serves to translate between sparse representations in CA3 and more overlapping representations in EC (for more discussion of this issue, see McClelland & Goddard, 1996; O'Reilly et al., 1998).

At test, 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 recurrents. 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.

To apply the hippocampal model to recognition, we exploited 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 exploited the fact that lures sometimes trigger recall of information that mismatches the recall cue. Thus, mismatch between recalled information and the test probe tends to indicate that an item was not studied.

For each test item, we generated 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 (a feature is counted as recalled if the unit corresponding to that feature in EC_out shows activity $> .9$); *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, one 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 one studied this particular item. 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.

The only place where we deviate from using *match − mismatch* in this article is in our related-lure simulations, where distractors are related to particular studied items but studied items are reasonably distinct from one another. In this situation, we use the recall-to-reject rule that places a stronger weight on mismatching recall. The *Simulation 3: YN Related-Lure Simulations* section of the article, below, contains a detailed account of our reasons for using recall-to-reject and the implications of using this rule in place of *match − mismatch*.

## Simulation Methodology

Our initial simulations involved a side-by-side comparison of the cortical and hippocampal networks receiving the exact same input patterns. This method allowed us to analytically characterize differences in how these networks responded to stimuli. A shortcoming of this side-by-side approach is that we could not explore direct interactions between the two

systems. To remedy this shortcoming, we also present simulations using a combined model where the cortical and hippocampal networks are connected in serial (such that the cortical regions involved in computing stimulus familiarity serve as the input to the hippocampal network)—this arrangement more accurately reflects how cortex and hippocampus are arranged in the brain (see Figure 1).

### *Basic Method*

In our recognition simulations, for each simulated participant we rerandomized the connectivity patterns for partial projections (e.g., if the specified amount of connectivity between Layer X and Layer Y was 25%, then each unit in Layer Y was linked at random to 25% of the units in Layer X), and we initialized the network weights to random values (weight values in the cortical model were sampled from a uniform distribution with mean = .5 and range = .25; see Appendix B for weight initialization parameters for different parts of the hippocampal model). The purpose of this randomization was to ensure that units in MTLC and the hippocampus would develop specialized receptive fields (i.e., they would be more strongly activated by some input features than others)—this kind of specialization is necessary for competitive learning.

After randomization was complete, 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 C). In our basic-parameter simulations, we used a 20-item study list (10 target items plus 10 interference items that were not tested), and the average amount of overlap between items was 20%—20% overlap between items was achieved by starting with a 24-slot prototype pattern and then generating items by randomly selecting a feature value for 16 randomly selected slots (note that each item overlapped at least 33% with the prototype but, on average, items overlapped 20% with each other).

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 (YN) and FC recognition tests. We matched performance in this way to alleviate concerns that differential effects of manipulations on hippocampal recall and MTLC familiarity might be 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.

### *Simulating YN and FC Testing*

To simulate YN recognition performance, items were presented one at a time at test, and we recorded the familiarity score (in the cortical model) and recall score (in the hippocampal model) triggered by each item. For the cortical model, we set an unbiased criterion for each simulated participant by computing the average familiarity scores associated with studied and lure items, respectively, and then placing the familiarity criterion exactly between the studied and lure means. All items triggering familiarity scores above this criterion were called *old*.

For the hippocampal model, we took a different approach to criterion setting; as discussed in *Simulation 2: Nature of the Underlying Distributions*, below, it is possible to set a high recall criterion that is sometimes crossed by studied items but never crossed by lures. We assumed that participants would be aware of this fact (i.e., that high amounts of recall are especially diagnostic of having studied an item) and set a recall criterion that was high enough to avoid false recognition. Accordingly, in our basic-parameter simulations, we used a fixed, relatively high criterion for calling items *old* (recall $\geq .40$). This value was chosen because—assuming other parameters were set to their basic values—it was sometimes exceeded by studied items but never by lures (unless lures were constructed to be similar to specific studied items; see *Simulation 3: YN Related-Lure Simulations*, below, for more details).

For both models, we used $d'$ (computed on individual participants' hit and false-alarm rates) to index YN recognition sensitivity.[1]

Our method for simulating FC testing was straightforward: We presented the two test alternatives one at a time. For the cortical model, we recorded the *act_win* score associated with each test alternative and selected the item with the higher score. For the hippocampal model, we recorded the *match* − *mismatch* score associated with each test alternative and selected the item with the higher score. For both models, if there was a tie, one of the two test alternatives was selected at random.

All of the simulation results reported in the text of this article are significant at $p < .001$. In graphs of simulation results (starting with *Simulation 3: YN Related-Lure Simulations*, below), error bars indicate the standard error of the mean computed across simulated participants. 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).

## Part 1: Basic Network Properties

Simulations reported in this section addressed 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*

Many of the differences between hippocampally and cortically driven recognition in our model arise from the fact that the hippocampal network exhibits more pattern separation than the cortical network. To document the two networks' pattern-separation abilities, we ran simulations where we manipulated the amount of overlap between paired input patterns. The first item in each pair was presented at study, and the second item was presented at test. For each pair, we measured the resulting amount of overlap in region CA3 of the hippocampal model and in the hidden (MTLC) layer of the cortical model. Pattern separation is evident when overlap between the internal representations of paired items (in CA3 or MTLC) is less than the amount of input overlap.

The results of these simulations (see Figure 5) confirm that, although both networks show some pattern separation, the amount of pattern separation is larger in the hippocampal model. They also show that the hippocampus's 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).

### *Simulation 2: Nature of the Underlying Distributions*

One way to characterize how cortical and hippocampal contributions to recognition differ is to plot the distributions of these

[1] $d' = z(H) - z(F)$, where $z$ is the inverse of the normal distribution function, $H$ is the hit rate, and $F$ is the false-alarm rate. To avoid problems with $d'$ being undefined when hit or false-alarm rates equalled 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 $n$ is the number of *old* responses, $N$ is the total number of items, and $P$ is the corrected-percentage *old* value.

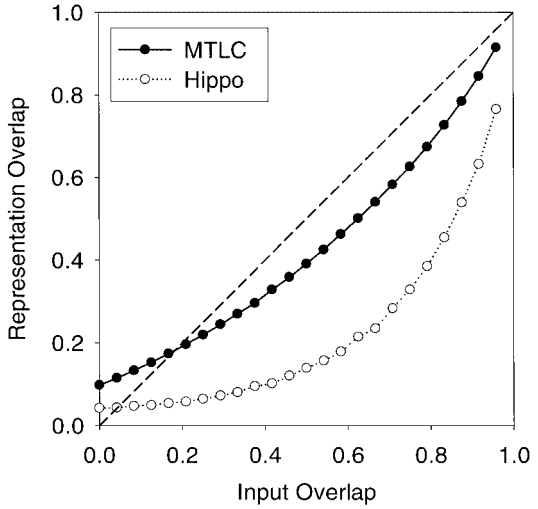


*Figure 5.* Results of simulations exploring pattern separation in the hippocampal and cortical models. In these simulations, we created pairs of items and manipulated the amount of overlap between paired items. The graph plots the amount of input-layer overlap for paired items versus (a) CA3 overlap in the hippocampal model and (b) medial temporal lobe cortex (MTLC) overlap in the cortical model. All points below the diagonal (dashed line) indicate pattern separation (i.e., representational overlap $<$ input overlap). The hippocampal model shows a strong tendency toward pattern separation (CA3 overlap $<<$ input overlap); the cortical model shows a smaller tendency toward pattern separation (MTLC overlap is slightly less than input overlap). Hippo = hippocampus.

signals for studied items and lures. Given 20% average overlap between input patterns, the MTLC familiarity distributions for studied items and lures are Gaussian and overlap strongly (see Figure 6A)—this is consistent with the idea, expressed by Yonelinas, Dobbins, Szymanski, Dhaliwal, and King (1996) and many others (e.g., Hintzman, 1988), that familiarity is well described by standard (Gaussian) signal-detection theory. In contrast, the hippocampal recall distributions (see Figure 6B) do not adhere to a simple Gaussian model. The bulk of the lure recall distribution is located at the zero recall point, although some lures trigger above-zero recall. The studied recall distribution is bimodal, and crucially, it extends further to the right than the lure recall distribution, so there are some (high) recall scores that are sometimes triggered by studied items but never by lures. Thus, high levels of recall are highly diagnostic—for these parameters, if an item triggers a recall score of .2 or greater, one can be completely sure that it was studied.

The low overall level of lure recall in this simulation can be attributed to hippocampal pattern separation. Because of pattern separation, the CA3 representations of lures do not overlap strongly with the CA3 representations of studied items. Because the CA3 units activated by lures (for the most part) were not activated at study, these units do not possess strong links to CA1; as such, activity does not spread from CA3 to CA1, and recall does not occur.

The studied recall distribution is bimodal because of nonlinear attractor dynamics in the hippocampus. If a studied test cue accesses a sufficiently large number of strengthened weights, it triggers pattern completion: Positive feedback effects (e.g., in the CA3 recurrents) result in strong reactivation of the CA3 and CA1 units that were activated at study, 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. These items only weakly activate CA3 and are poorly recalled, thereby accounting for the extra peak at recall = 0.

Increasing the average amount of overlap between items reduces the diagnosticity of the hippocampal recall signal. When the amount of overlap between input patterns is high (e.g., 40.5% instead of 20%), both studied items and lures trigger large amounts of recall, such that the studied and lure recall distributions are roughly Gaussian and overlap extensively (see Figure 7).

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 5), 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

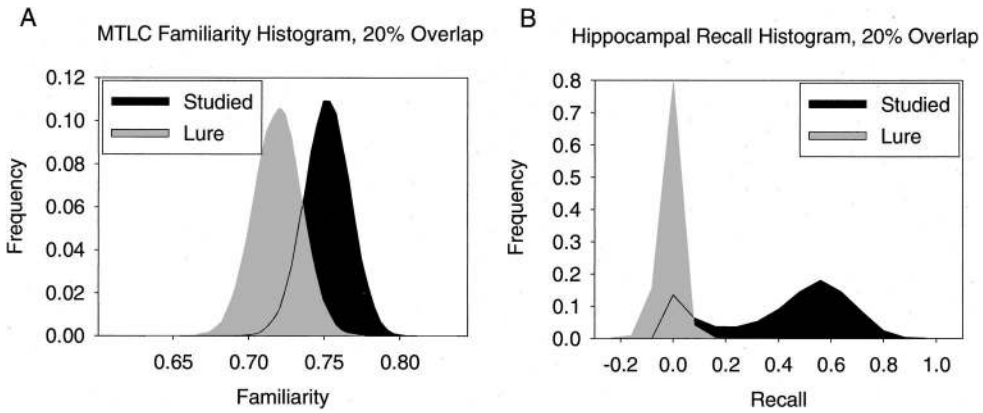


*Figure 6.* A: Histogram of the studied and lure medial temporal lobe cortex (MTLC) familiarity distributions for 20% average overlap. B: Histogram of the studied and lure hippocampal recall distributions for 20% average overlap.

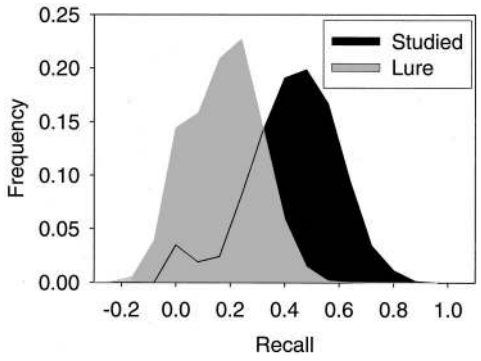


*Figure 7.* Histogram of the studied and lure hippocampal recall distributions for 40.5% average overlap.

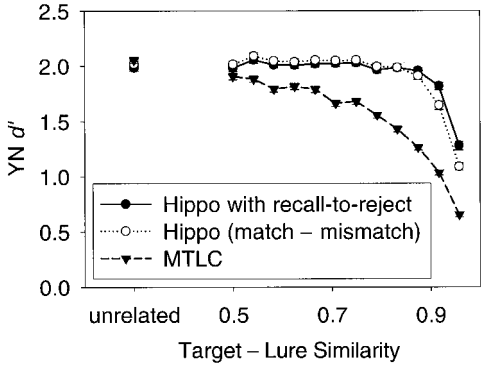


*Figure 9.* Yes–no (YN) recognition sensitivity ($d'$) in the two models, as a function of target–lure similarity. Target–lure similarity was operationalized as the proportion of input features shared by targets and corresponding lures; note that the average level of overlap between studied (target) 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. The figure also shows that hippocampal performance for related lures is better when a recall-to-reject rule is used instead of match − mismatch. Hippo = hippocampus; MTLC = medial temporal lobe cortex.

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 in response to both studied items and lures. Figure 8 shows that increasing overlap increases the probability that prototypical features of studied items and lures will be recalled and reduces the probability that item-specific features of studied items will be recalled (in part, because the hippocampus intrudes prototypical features in place of these item-specific features).

In summary, the results presented here show that hippocampal recall has two modes of operation: When input patterns have low to moderate average overlap, high levels of matching recall are highly diagnostic—studied items sometimes trigger strong recall (of item-specific and prototype features), but lures trigger virtually no recall. 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.

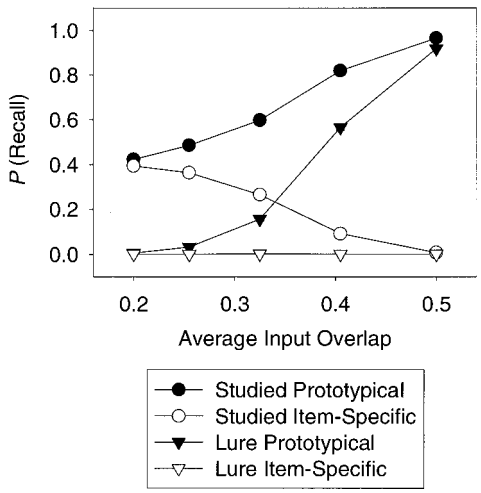


*Figure 8.* Plot of the probability that item-specific and prototypical features of studied items and lures will be recalled, as a function of overlap. As overlap increases, the amount of prototype recall triggered by studied items and lures increases, and the amount of item-specific recall triggered by studied items decreases.

### *Simulation 3: YN Related-Lure Simulations*

The strengths of the hippocampal model are most evident on YN related-lure recognition tests, where lures are similar to studied items but studied items are dissimilar to one another. In this section, we show how the hippocampal model outperforms the cortical model on these tests because of its superior pattern-separation and pattern-completion abilities.

*Method.* To simulate the related-lure paradigm, we first created studied-item patterns with 20% average overlap between items. Then, for each studied (target) item, we created a related-lure item by taking the studied item and flipping a prespecified number of slots; to vary target–lure similarity, we varied the number of slots that we flipped to generate lures (less flipping resulting in more overlap). For comparison, we also ran simulations with unrelated lures that were sampled from the same item pool as studied items.

In the related-lure simulations presented here (and later in the article), we used a recall-to-reject hippocampal decision rule instead of our standard *match − mismatch* rule. According to this rule, items that trigger any mismatch are given a *new* response, otherwise the decision is based on match. We used this rule for two reasons: First, there is extensive empirical evidence that, when lures are similar to studied items but studied items are unrelated to one another, participants use recall-to-reject (see, e.g., Rotello & Heit, 2000; Rotello, Macmillan, & Van Tassel, 2000; Yonelinas, 1997; but see Rotello & Heit, 1999, for a contrasting view). Second, it is computationally sensible to use recall-to-reject—we show that the presence of mismatching recall in this paradigm is highly diagnostic of an item being nonstudied.

*Results.* Figure 9 shows the results of our related-lure simulations: Recognition performance based on MTLC familiarity gets steadily worse as lures become increasingly similar to studied items; in contrast, recognition based on hippocampal recall is relatively robust to the lure-similarity manipulation. Figure 9 also shows that hippocampal performance is somewhat better for re-

lated lures when the recall-to-reject rule is used (as opposed to *match − mismatch*).

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 occurred 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 is greater than 60%, and lures do not start to trigger above-criterion (i.e., $> .40$) recall until target–lure similarity is greater than 80%. This occurs because of hippocampal pattern separation—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 lures sometimes trigger pattern completion of the corresponding studied item and can subsequently be rejected based on mismatch between recalled information and the test cue. Figure 10 illustrates the point (mentioned earlier) that when lures resemble studied items but studied items are not related to one another, mismatching recall is highly diagnostic—studied items virtually never trigger mismatching recall, but lures sometimes do. As such, it makes sense to use a rule (like recall-to-reject) that assigns a very high weight to mismatching recall.

Figure 10 also shows that mismatching recall triggered by lures increases substantially with increasing target–lure similarity. This increase in mismatching recall helps offset, to some degree, increased matching recall triggered by related lures. With recall-to-reject, the only way that lures can trigger an *old* response is if they trigger a large amount of matching recall but no mismatching recall. The odds of this happening are very low.

In summary, these simulations demonstrate that both networks can support good performance on YN recognition tests with lures that are unrelated to studied items. When the networks are challenged by boosting target–lure similarity, performance in both networks suffers; however, the hippocampus is more robust to this manipulation than cortex. As such, the model predicts that recognition discrimination based on hippocampal recall should be better than discrimination based on MTLC familiarity on YN tests with related lures. This prediction is consistent with the view, expressed in several empirical studies, that recall is especially important for discriminating between studied items and very similar distractors (see, e.g., Hintzman, Curran, & Oppy, 1992; Rotello et al., 2000).

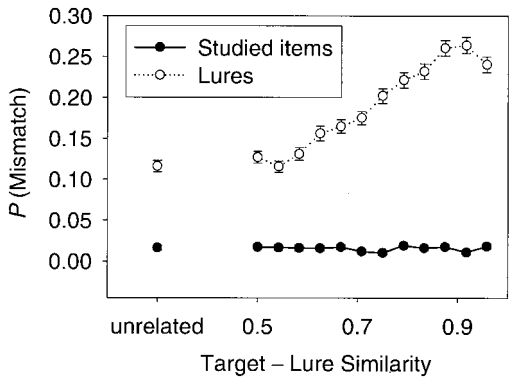


*Figure 10.* Plot of the probability that lures and studied items will trigger mismatching recall, as a function of target–lure similarity. This probability is close to floor for studied items; the probability that lures will trigger mismatching recall increases with increasing target–lure similarity.

### *Sources of Variability*

The final issue that we need to address in this Part 1: Basic Network Properties section is variability. 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 show later—different forms of variability have different implications for recognition performance.

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. 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).

Sampling variability arises because, at the beginning of each simulation, weight strengths and connectivity patterns are set randomly. As discussed earlier, this randomness helps units in MTLC and the hippocampus form specialized representations of the input space. It also has the consequence that, by chance, some input features are sampled better (by units further downstream) than other input features.

An important property of sampling variability is that it decreases as network size increases. Intuitively, as the number of units and connections increases, the odds that any one input pattern will be undersampled relative to another decreases. We conducted simulations to explore this issue, and the results are very clear: As we increase network size, variability in MTLC-familiarity and hippocampal-recall scores steadily decreases, and $d'$ scores steadily increase.

In a network scaled to the approximate size of the human brain, sampling variability would likely be negligible. Therefore, we conclude that other forms of variability must be at play in the human brain; we briefly describe some other sources of variability below.

### *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 participants' 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 related lures in tandem. That is, encoding fluctuations that boost the

memory signal triggered by a studied item 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; in small networks where sampling variability is the dominant source of variance, noise associated with sampling of nonshared (discriminative) features of overlapping stimuli counteracts much of the shared variability in memory scores triggered by these items. We revisit this issue later, when we explore how lure relatedness interacts with test format (*Simulation 4: Lure-Relatedness and Test-Format Interactions*, below).

Another source of variability in recall and familiarity scores is variability in preexperimental 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 preexperimental presentation frequency affects recognition memory, see Dennis and Humphreys (2001). Variability in preexperimental exposure (like encoding variability, but unlike sampling variability) affects studied items and related lures in tandem.

Finally, in addition to variability in how much test items overlap with preexperimental 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.

### *Sources of Variability: Summary*

In summary, the basic model (as described above) is strongly influenced by sampling variability and lacks other plausible sources of variability such as encoding variability. Given that sampling variability is not likely to be a factor in human recognition-memory performance, one might conclude that this source of variability should be eliminated and other sources incorporated. 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).

Because of these limitations, we refrain in this article from making strong predictions about how manipulations affect variance. Nevertheless, we can still use the basic model to explain many phenomena that do not depend on the exact source of variability. Also, it is relatively straightforward to supplement the basic model with other forms of variability on an as-needed basis, and we do this to make some important points in *Simulation 4: Lure-Relatedness and Test-Format Interactions*, below.

## Part 2: Applications to Behavioral Phenomena

The simulations in this part of the article 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 present data that speak to the model's predictions.

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

As we showed in Figure 9, the model predicts that the hippocampus should outperform cortex on standard YN tests where participants have to discriminate between studied items and related lures. In this section, we show how giving participants a forced choice between studied items and corresponding related lures benefits performance in the cortical model but not the hippocampal model, thereby mitigating the hippocampal advantage.

### *FC Testing and Covariance*

In an FC test, participants 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 are reliably 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 are reliable.

The reliability of the familiarity difference depends on where variability comes from in the model. As discussed in the previous section, some kinds of variability (e.g., differences in encoding strength and preexperimental 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 sampling variability predominates, the studied-lure familiarity difference is somewhat 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). The variance of the studied-lure familiarity difference is given by the following equation:

$$Var(S - L) = Var(S) + Var(L) - 2 \times Cov(S, L), \qquad (2)$$

where $S$ represents familiarity of studied items, $L$ that of lures, and *Var* is variance and *Cov* covariance between $S$ and $L$. Equation 2 shows that increasing covariance reduces the variance of the $S$–$L$ familiarity difference, which in turn boosts FC performance.

*FC simulations using the basic cortical model.* Our first concern was to assess how much FC testing with corresponding related lures benefits performance in our basic cortical model. Toward this end, we ran simulations using a paradigm introduced by Hintzman (1988); in these simulations, 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 noncorresponding 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 noncorresponding lures.

As shown in Figure 11, FC performance is higher with corresponding related lures than with noncorresponding 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 computed the following ratio:

$$R = (2 \times Cov(S, L))/(Var(S) + Var(L)). \quad (3)$$

When $R = 1$, covariance completely offsets studied and lure variance, and the studied-lure familiarity difference is completely reliable (i.e., variance = 0); $R = 0$ means that there is no covariance. For target–lure similarity = .92, the covariance ratio $R = .27$ in the corresponding condition, and $R = -.01$ in the noncorresponding condition. Thus, the model exhibits roughly one third the maximal level of covariance possible.

In summary, although the basic model qualitatively exhibits an 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.

*Cortical and hippocampal simulations with encoding variability.* Next, we wanted to explore a more realistic scenario in which the contribution of sampling variability to overall variability was small relative to other forms of variability (such as encoding variability) that affect studied items and corresponding lures in tandem. Increasing the relative contribution of encoding variability should increase covariance and thereby increase the extent to which the cortical model benefits from use of an FC-corresponding test. We were also interested in how test format affects the hippocampal model's ability to discriminate between studied items and related lures (when encoding variability is high). To address these issues, we ran simulations with added encoding variability in both the cortical and hippocampal models where we manipulated test format (FC corresponding vs. FC noncorresponding vs. YN) and lure relatedness.

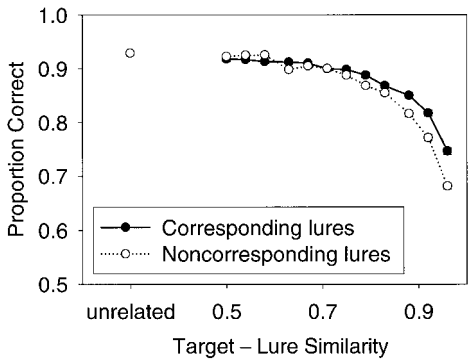


*Figure 11.* Forced-choice (FC) accuracy in the cortical model as a function of target–lure similarity, using corresponding and noncorresponding FC testing. For high levels of target–lure similarity, FC performance is slightly better with corresponding lures than with noncorresponding lures.

*Method.* We added encoding variability 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 did not achieve the desired result; the influence of encoding variability was still too small relative to sampling variability, and overall performance levels with added encoding variability were unacceptably low. To boost the relative impact of encoding variability (and overall performance), we also increased the learning rate in both models to 3 times its usual value. 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 being asked here. As with the hippocampal related-lure simulations presented earlier, the hippocampal simulations presented here used a recall-to-reject decision rule. We applied this rule to FC testing in the following manner: If one item triggered mismatching recall but the other item did not, the second item was selected; otherwise, the item triggering a higher match–mismatch recall score was selected.

*Cortical FC results.* As expected, the corresponding versus noncorresponding difference for the cortical model is much larger when encoding variability is present (see Figure 12A) than when encoding variability is absent (see Figure 11). Computing the average covariance/variance ratio for the .92 target–lure overlap condition shows that $R = .62$ for corresponding lures versus $R = -.05$ for noncorresponding lures. This is more than double the covariance in the basic model (.62 vs. .27) and confirms our intuition that decreasing the contribution of sampling variability relative to encoding variability would increase covariance and boost performance in the FC-corresponding condition.

*Hippocampal FC results.* In contrast to the cortical model results, FC-corresponding and FC-noncorresponding performance are almost identical in the hippocampal model (see Figure 12A). 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 happened on trials where the studied item was not recalled—on these trials, participants can still respond correctly if the lure triggers mismatching recall (and is rejected on this basis). The key insight is that studied recall and lure misrecall are independent when noncorresponding lures are used (in effect, participants 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 participants 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.

Because the cortical model benefits from FC-corresponding (vs. noncorresponding) testing but the hippocampal model does not, the performance of the cortical model relative to the hippocampal model is better in this condition.

*YN results.* The results of our YN simulations with encoding variability (Figure 12B) are identical to the results of our earlier YN–related-lure simulations. As before, we found that the hip-

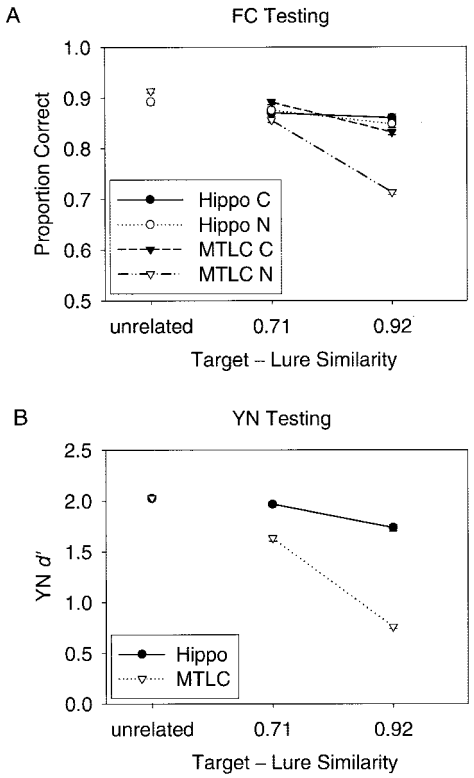


*Figure 12.* Cortical and hippocampal related-lure simulations incorporating strong encoding variability. A: Results of forced-choice (FC) simulations. When encoding variability is present, the cortical model benefits very strongly from use of corresponding versus noncorresponding lures (more so than in our simulations without encoding variability). In contrast, the hippocampal model (using recall-to-reject) performs equally well with corresponding and noncorresponding lures. B: Results of yes–no (YN) simulations with the same parameters. As in our previous related-lure simulations (see Figure 9), the cortical model is severely impaired relative to the hippocampal model on these tests. Hippo = hippocampus; MTLC = medial temporal lobe cortex; C = corresponding lures; N = noncorresponding lures.

pocampal model is much better than the cortical model at discriminating between studied items and related lures on YN tests.

### *Tests of the Model's Predictions*

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 results shown in Figure 12, patients should be impaired on YN recognition tests with related lures, but they should perform relatively well on FC-corresponding tests with related lures and on tests with unrelated lures (regardless of test format).

To test this prediction, we collaborated with Andrew Mayes and Juliet Holdstock to test patient Y.R., who suffered focal hippocampal damage sparing surrounding MTLC regions (for details of the etiology and extent of Y.R.'s lesion, see Holdstock et al., 2002). Y.R. is severely impaired at recalling specific details—thus, Y.R. has to rely almost exclusively on MTLC familiarity when making recognition judgments. Holdstock et al. (2002) developed YN and FC tests with highly related lures that were closely matched for difficulty and administered these tests to patient Y.R. and her controls. Figure 13 shows sample stimuli from this experiment. Results from this experiment can be compared with 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., 2002); we refer to these tests as the *YN–low-relatedness* and *FC–low-relatedness tests*, respectively.

Figure 14 shows that, exactly as we predicted, Y.R. 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 (Y.R. actually performed slightly better than the control mean on this test). This pattern cannot be explained in terms of difficulty confounds (i.e., Y.R. 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 14 also shows that Y.R. was, on average, unimpaired on YN–low-relatedness and FC–low-relatedness tests. Y.R. performed worse on the YN test with highly related lures than on any of the 15 YN–low-relatedness tests. This difference cannot be attributed to the YN–low-relatedness tests being easier than the YN test with highly related lures: Y.R. 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 8 tests, her mean $z$ score was 0.04 ($SD$ = 0.49; minimum = −0.54; maximum = 0.65; J. Holdstock, personal communication

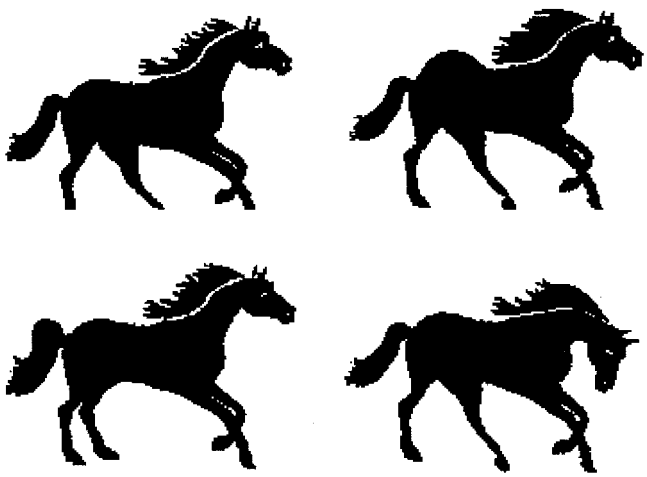

*Figure 13.* Sample stimuli from the Holdstock et al. (2002) related-lure experiment. Participants studied pictures of objects (e.g., the horse shown in the upper left). At test, participants had to discriminate studied pictures from three highly related lures (e.g., the horses shown in the upper right, lower left, and lower right). From "Under What Conditions Is Recognition Spared Relative to Recall After Selective Hippocampal Damage in Humans?" by J. S. Holdstock et al., 2002, *Hippocampus, 12,* p. 344. Copyright 2002 by Wiley. Reprinted with permission.

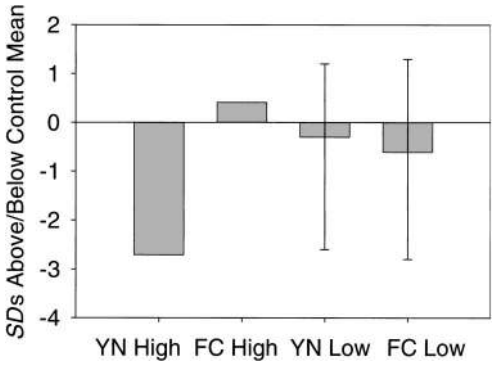


*Figure 14.* Performance of patient Y.R. relative to controls on matched yes–no (YN) and forced-choice (FC) corresponding tests with highly related (High) lures; the graph also plots Y.R.'s average performance on 15 YN tests and 25 FC tests with less strongly related (Low) lures. Y.R.'s scores are plotted in terms of number of standard deviations (*SD*s) 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 Y.R. (across the 15 YN tests and the 25 FC tests, respectively). Y.R. was significantly impaired relative to controls on the YN test with highly related lures (i.e., her score was $> 1.96$ *SD*s below the control mean), but Y.R. performed slightly better than controls on the FC test with highly related lures. Y.R. was not significantly impaired, on average, on the tests that used less strongly related lures.

December 15, 2000). We have yet to test the model's prediction regarding use of FC-corresponding versus FC-noncorresponding tests with related lures; on the basis of the results shown in Figure 12, the model predicts that Y.R. will be more strongly impaired on FC tests with noncorresponding (vs. corresponding) related lures.

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

In this section, we explore the two networks' performance on associative-recognition tests. On these tests, participants have to discriminate between studied pairs of stimuli (A–B, C–D) 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. The hippocampus's 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 (see, e.g., Rolls, 1989; Rudy & Sutherland, 1995; Squire, 1992b; Teyler & Discenna, 1986). In 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.

Associative recognition can be viewed as a special case of the related-lure paradigm described earlier. As such, we would expect the hippocampus to outperform cortex on YN associative-recognition tests because of its superior pattern-separation abilities and its ability to reject similar lures based on mismatching recall.

#### *Associative Recognition*

*Method.* 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 nonstudied item), and *novel* lures (generated by pairing two nonstudied items). Our initial simulations used a YN test format.

*YN results.* As expected, the hippocampal model outperforms the cortical model on this YN associative-recognition test (see Figure 15). We also found that cortical performance 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*WTA algorithm). The *k*WTA 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 sensitive only 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. For additional computational support for the notion that cortex should encode low-order conjunctive representations, see O'Reilly and Busby (2002).

#### *Effects of Test Format*

In *Simulation 4: Lure-Relatedness and Test-Format Interactions*, above, we showed how giving the models a forced choice between studied items and corresponding related lures mitigates the hippocampal advantage for related lures. Analogously, giving the models a forced choice between overlapping studied pairs and lures (FC-OLAP testing: study A–B, C–D; test A–B vs. A–D) mitigates the hippocampal advantage for associative recognition. In both cases, performance suffers because the hippocompal model

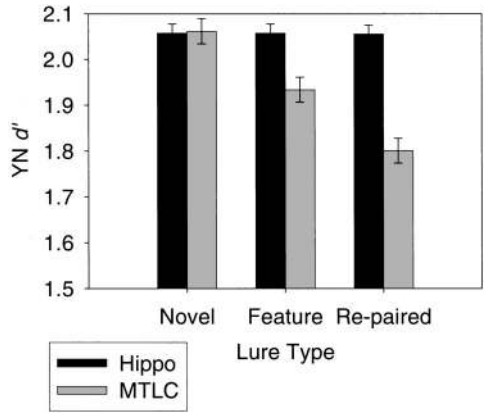


*Figure 15.* Results of yes–no (YN) associative-recognition simulations in the cortical (MTLC) and hippocampal (Hippo) models. With parameters that yield matched performance for unrelated (novel) lures, cortex is impaired relative to the hippocampus at associative recognition; nonetheless, cortex performs well above chance on the associative-recognition tests.

does not get an extra chance to respond correctly (via recall-to-reject) when studied recall fails.

Typically, FC-OLAP tests are structured in a way that emphasizes the shared item: Participants are asked, "Which of these items was paired with A: B or D?" This encourages participants to use a strategy where they cue with the shared item (A) and select the choice (B or D) that best matches retrieved information. The problem with this algorithm is that success or failure depends entirely on whether the shared cue (A) triggers recall; if A fails to trigger recall of B, participants are forced to guess. In contrast, on FC associative-recognition tests with nonoverlapping choices (FC-NOLAP tests: study A–B, C–D, E–F; test A–B vs. C–F), participants have multiple, independent chances to respond correctly; even if A does not trigger recall of B, participants can still respond correctly if they recall that C was paired with D (not F).

To demonstrate how recall-to-reject differentially benefits FC-NOLAP performance (relative to FC-OLAP performance), we ran FC-NOLAP and FC-OLAP simulations in the hippocampal model using a recall-to-reject rule. For comparison purposes, we ran another set of simulations where decisions were based purely on the amount of matching recall.

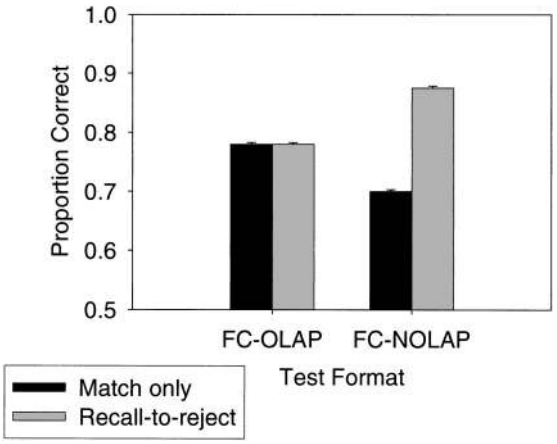


*Figure 16.* Associative-recognition performance in the hippocampal model, as a function of recall decision rule (match only vs. recall-to-reject) and test format (forced-choice with overlapping choices [FC-OLAP] vs. forced-choice with nonoverlapping choices [FC-NOLAP]). FC-NOLAP performance benefits from use of the recall-to-reject rule, but FC-OLAP performance does not benefit at all. When the recall-to-reject rule is used, FC-NOLAP performance is better than FC-OLAP performance.

### *Test Format Simulation*

*Method.* These simulations used a cued recall algorithm where, for each test pair (e.g., A–B), we cued with the first pair item and measured how well recalled information matched the second pair item. On FC-OLAP tests, we cued with the shared pair item (A) for both test alternatives (A–B vs. A–D). On FC-NOLAP tests, we cued with the first item of each test alternative (A from A–B and C from C–D). 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, and we increased the activation criterion for counting a feature as recalled (from .90 to .95) to compensate for the fact that the output of the model was less clean with partial cues.

*Results.* As expected, FC-NOLAP performance is higher in the recall-to-reject condition (vs. the match-only condition), but FC-OLAP performance does not benefit at all (see Figure 16). Because of this differential benefit, FC-NOLAP performance is better overall than FC-OLAP performance in the recall-to-reject condition. 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 participants were using recall of studied pairs to reject lures and that participants had more unique (independent) chances to recall useful information in the FC-NOLAP condition.

### *Tests of the Model's Predictions*

The implications of the above 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. Kroll, Knight, Metcalfe, Wolf, and Tulving (1996) studied associative-recognition performance in a patient with bilateral hippocampal damage (caused by anoxia), as well as in other patients with less focal lesions. In Experiment 1 of the Kroll et al. study, participants studied two-syllable words (e.g., *barter*, *valley*) and had to discriminate between studied words and words created by recombining studied words (e.g., *barley*). Results from the patient with bilateral hippocampal damage, as well as control data, are plotted in Figure 17. In keeping with the model's predictions, the patient showed impaired YN associative-recognition performance, 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, participants did not necessarily have to form a new conjunctive representation to solve this task.

Two studies (Mayes et al., 2001; Vargha-Khadem et al., 1997) have examined how well patients with focal hippocampal damage perform on FC-OLAP tests where participants were cued with one pair item and had 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 both of these studies, the hippocampally lesioned patients were unimpaired. This is consistent with the model's prediction that patients should perform relatively well, compared with controls, on FC-OLAP tests. Furthermore, despite the patients' having hippocampal lesions, their excellent performance on these tests, 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

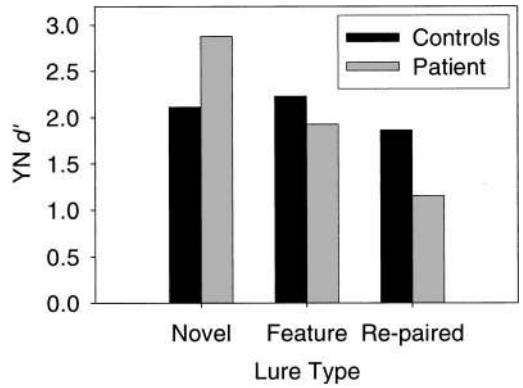


*Figure 17.* Associative recognition in a patient with bilateral hippocampal damage (from Kroll, Knight, Metcalfe, Wolf, & Tulving, 1996, Experiment 1); $d'$ scores for the patient and controls were computed based on average hit and false-alarm rates published in Table 3 of Kroll et al. The patient performed 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 recombining studied stimuli. The pattern reported here is qualitatively consistent with the model's predictions as shown in Figure 15, above. YN = yes–no.

study exposure. One caveat is that, although MTLC appears capable of supporting good associative-recognition performance when the to-be-associated stimuli are the same kind (e.g., words), it performs less well when the to-be-associated stimuli are of different kinds (e.g., objects and locations; Holdstock et al., 2002; Vargha-Khadem et al., 1997). Holdstock et al. (2002) argued that MTLC familiarity cannot support object–location associative recognition because object and location information do not converge fully in MTLC.

As a final note, although the studies discussed above found patterns of spared and impaired recognition performance after focal hippocampal damage that are consistent with the model's predictions, it is important to keep in mind that many studies have found an across-the-board impairment in declarative-memory tasks following hippocampal damage. For example, Stark and Squire (2002) found that patients with hippocampal damage were impaired to a roughly equal extent on tests with novel versus re-paired lures. We address the question of why some hippocampal patients show across-the-board versus selective deficits in *Simulation 8: Lesion Effects in the Combined Model*, below.

### *Simulation 6: Interference and List Strength*

We now turn to the fundamental issue of interference: How does studying an item affect recognition of other, previously studied items? In this section, we first review general principles of interference in networks like ours that incorporate Hebbian LTP and LTD. We then show how a list-strength interference manipulation (described in detail below) differentially affects discrimination based on hippocampal recall versus MTLC familiarity.

#### *General Principles of Interference in Our Models*

At the most general level, interference occurs in our models whenever different input patterns have overlapping internal representations. In this situation, studying a new pattern tunes the overlapping units so they are more sensitive to the new pattern and less sensitive to the unique (discriminative) features of other patterns.

Figure 18 is a simple illustration of this tuning process. It shows a network with a single hidden unit that receives input from five input units. Initially, the hidden unit is activated to a roughly equal extent by two different input patterns, A and B. Studying Pattern A has two effects: Hebbian LTP increases weights to active input features, and Hebbian LTD decreases weights to inactive input features. These changes bolster the extent to which Pattern A activates the hidden unit. The effects of learning on responding to Pattern B are more complex: LTP boosts weights to features that are shared by Patterns A and B, but LTD reduces weights to features that are unique to Pattern B.

If one trains a network of this type on a large number of overlapping patterns (e.g., several pictures of fish), the network becomes 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 less and less responsive to discriminative features of individual stimuli (e.g., the fact that one fish has a large green striped dorsal fin). In the long run, this latter effect is harmful to recognition performance—if the network's sensitivity to the unique, discriminative features of studied items and lures hits floor, then the network is not able to respond differentially to studied items and lures at test. However, in the absence of floor effects, the extent to which recognition is harmed 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 Simulations*

We begin our exploration of interference by simulating how list strength affects recognition in the two models; specifically, how does strengthening some list items affect recognition of other (nonstrengthened) list items (Ratcliff, Clark, & Shiffrin, 1990)?

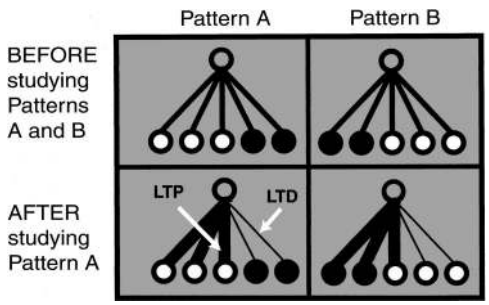


*Figure 18.* Illustration of how Hebbian learning causes interference in our models. Initially (top two squares), the hidden unit responds equally to Patterns A and B. The effects of studying Pattern A are shown in the bottom two squares. Studying Pattern A boosts weights to features that are part of Pattern A (including features that are shared with Pattern B) and reduces weights to features that are not part of Pattern A. These changes result in a net increase in responding to Pattern A and a net decrease in responding to Pattern B. LTP = long-term potentiation; LTD = long-term depression.

*Method.* In our list-strength simulations, we compared two conditions: a weak-interference condition and a strong-interference condition. In the weak-interference condition, the model was given a study list consisting of target items presented once and interference items presented once. The strong-interference condition was the same except that interference items were strengthened at study by presenting them multiple times. In both conditions, the model had to discriminate between target items and nonstudied lures at test. If strengthening of interference items (in the strong-interference condition) impairs target versus lure discrimination relative to the weak-interference condition, this is an LSE. A simple diagram of the procedure is provided in Table 1.

The study list was comprised of 10 target items followed by 10 interference items. Interference-item strength was manipulated by increasing the learning rate for these items (from .01 to .02). 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 are on the next presentation, and greater activity leads to greater learning (according to the Hebb rule). 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.

*Results.* In the cortical network (see Figure 19A), there is no effect of list strength on recognition when input-pattern overlap is relatively low (up to .26), but the LSE is significant for higher levels of input overlap. In contrast, the hippocampal network shows a significant LSE for all levels of input overlap (see Figure 19B); the size of the hippocampal LSE increases with increasing overlap (except in the .5 overlap condition, where the LSE is compressed by floor effects). Figure 19C 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 exhibit a robust item-strength effect whereby memory for interference items is better in the strong-interference condition (e.g., for 20% input overlap, interference-item $d'$ increases from 2.13 to 3.22 in the hippocampal model; in the cortical model, $d'$ increases from 2.08 to 3.12), thereby confirming that our strengthening manipulation is 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.

Table 1
*List-Strength Procedure*

| Target items | | Interference items | | | |
|---|---|---|---|---|---|
| | | Weak interference | | | |
| bike | robot | apple | cat | tree | towel |
| | | Strong interference | | | |
| bike | robot | apple | cat | tree | towel |
| | | | cat | tree | towel |
| | | | cat | tree | towel |

*Note.* List-strength simulations compared weak-interference lists with strong-interference lists. In both conditions, the model had to discriminate between targets (e.g., bike) and nonstudied lures (e.g., coin) at test.

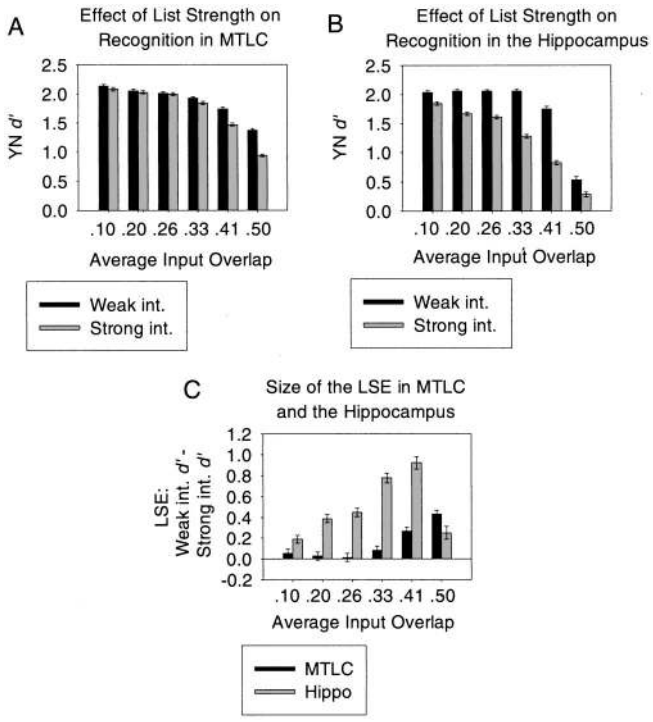


*Figure 19.* Results of list-strength simulations in the two models. A: Effect of list strength on recognition in medial temporal lobe cortex (MTLC). B: Effect of list strength on recognition in the hippocampus (Hippo). C: Size of the list-strength effect (LSE) in MTLC and the hippocampus; this panel replots 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 is a significant LSE in the hippocampal model but not in the cortical model; for higher levels of overlap, there is an LSE in both models. YN = yes–no; Strong int. = strong interference; Weak int. = weak interference.

### *Interference in the Hippocampal Model*

Understanding the hippocampal LSE is quite straightforward. Even though there is less overlap between representations in the hippocampus than in cortex, there is still some overlap (primarily in CA1, but also in CA3). These overlapping units cause interference—specifically, recall of discriminative features of studied items is impaired through Hebbian LTD occurring in the CA3–CA1 projection and (to a lesser extent) in projections coming into CA3. Importantly, for low to moderate levels of input-pattern overlap, the amount of recall triggered by lures is at floor, and therefore, it cannot decrease as a function of interference. Putting these two points together, the net effect of interference is to move the studied distribution downward toward the at-floor lure distribution, which increases the overlap between distributions and therefore impairs discriminability (see Figure 20).

### *Interference in the Cortical Model*

Next, we need to explain why an LSE is 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 actually 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.

Figure 21A shows how raw familiarity scores triggered by targets and lures change as a function of interference (for 20% overlap). Initially, both studied and lure familiarity decrease; this occurs because interference reduces weights to discriminative features of both studied items and lures. There is also an interaction whereby lure familiarty decreases more quickly than studied familiarity, so the studied–lure gap in familiarity increases slightly (see Figure 21B; note that although the increase is numerically small, it is highly significant).

However, with more interference, the studied–lure gap in familiarity starts to decrease. This occurs because weights to discriminative features of lures eventually approach floor (and—as such—cannot show any additional decrease as a function of interference). Also, raw familiarity scores start to increase; this occurs because, in addition to reducing weights to discriminative features, interference also boosts weights to shared-item features. At first, the former effect outweighs the latter, and there is a net decrease in familiarity. However, when weights to discriminative features approach floor, these weights no longer decrease enough to offset the increase in weights to shared features, and there is a net increase in familiarity.

For these parameters, list strength does not substantially affect the variance of the familiarity signal; variance increases numerically with increasing list strength, but the increase is extremely small (e.g., 10-times strengthening of the 10 interference items leads to a 3% increase in variance). However, we cannot rule out the possibility that adding other forms of variability to the model (and eliminating sampling variability; see the Sources of Variability section above) might alter the model's predictions about how list strength affects variance.

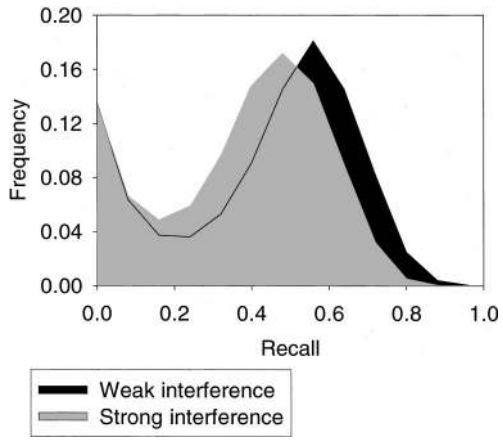


*Figure 20.* Studied-recall histograms for the strong- and weak-interference conditions, 20% overlap condition. Increasing list strength pushes the studied-recall distribution to the left (toward zero).

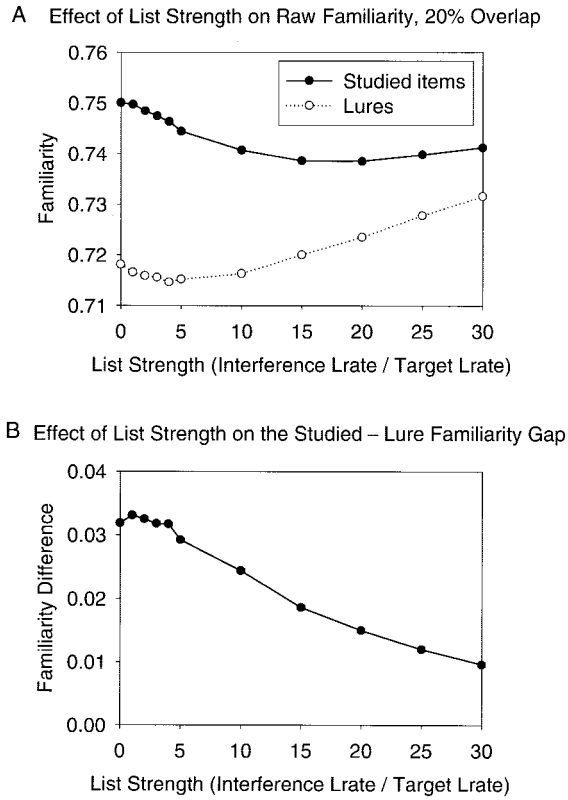


*Figure 21.* A: Plot of how target and lure familiarity are affected by list strength (with 20% input overlap). Initially, target and lure familiarity decrease; however, with enough interference, target and lure familiarity start to increase. B: Plot of the difference in target and lure familiarity, as a function of list strength; initially, the difference increases slightly, but then it decreases. Lrate = learning rate.

*Differentiation.* The finding that lure familiarity initially decreases faster than studied familiarity can be explained in terms of the principle of differentiation, which was first articulated by Shiffrin et al. (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.

In our model, differentiation is a simple consequence of Hebbian learning (as it is in McClelland & Chappell, 1998). As discussed above, Hebbian learning tunes 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; as such, studied items suffer less interference than lures. As an example of how studying an item pulls its representation away from other items, with 20% input overlap, the average amount of MTLC overlap between studied target items and interference items (expressed in terms of vector dot product) is .150, whereas the average overlap between lures and interference items is .154; this difference is highly significant.

*Boundary conditions on the null LSE.* It should be clear from the above explanation that we do not always expect a null LSE for

MTLC familiarity. With enough interference, the cortical 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 22, which plots target and lure familiarity as a function of list strength, for 40.5% input overlap.

### *Tests of the Model's Predictions*

The main novel prediction from our models is that (modulo the boundary conditions outlined above) recognition based on hippocampal recall should exhibit an LSE, whereas recognition based on MTLC familiarity should not.

Consistent with the hippocampal model's prediction, some studies have found an LSE for cued recall (see, e.g., Kahana, Rizzuto, & Schneider, 2003; Ratcliff et al., 1990), although not all studies that have looked for a cued-recall LSE have found one (see, e.g., Bauml, 1997). However, practically all published studies that have looked for an LSE for recognition have failed to find one (Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990; Ratcliff, Sheu, & Gronlund, 1992; Shiffrin, Huber, & Marinelli, 1995; Yonelinas, Hockley, & Murdock, 1992). 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: LSEs should be obtained for recognition tests and measures that load more heavily on the recall process. Norman (2002), tested this prediction in two distinct ways.

*Self-report measures.* In one experiment, Norman (2002) collected self-report measures of recall and familiarity—whenever a participant called an item *old*, he or she was asked whether he or she specifically remembered details from when the item was studied or whether the item just seemed familiar. To estimate recall-based discrimination, Norman plugged the probability of saying "remember" to studied items and lures into the formula for $d'$ and computed familiarity-based discrimination using the independence remember–know technique described in Jacoby, Yonelinas, and Jennings (1997).

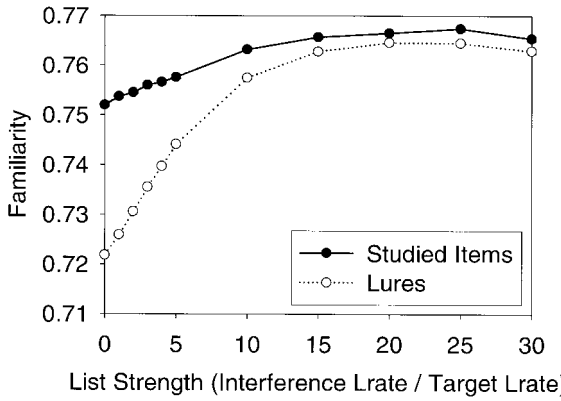


*Figure 22.* Plot of how target and lure familiarity are affected by list strength with 40.5% input overlap. When overlap is high, target and lure familiarity increase right from the start, and the target–lure familiarity gap monotonically decreases. Lrate = learning rate.

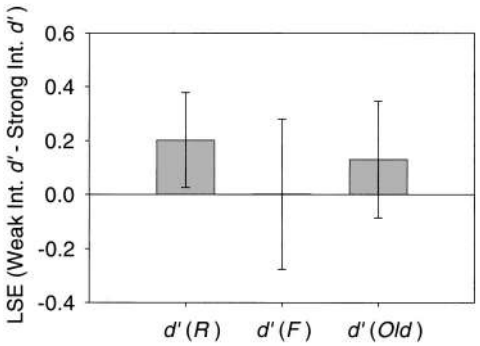


*Figure 23.* Plot of the size of the list-strength effect (LSE) for three dependent measures: $d'(R)$, recall-based discrimination; $d'(F)$, familiarity-based discrimination; and $d'(Old)$, discrimination computed based on *old/new* responses. Error bars indicate 95% confidence intervals. The LSE was significant for $d'(R)$ but not for $d'(F)$ or $d'(Old)$. Strong int. = strong interference; Weak int. = weak interference.

Norman (2002) found that the effect of list strength on *old/new* recognition sensitivity was nonsignificant in this experiment, replicating the null LSE obtained by Ratcliff et al. (1990). However, if one breaks recognition into its component processes, it is clear that list strength does affect performance (see Figure 23). As predicted, there was a significant LSE for discrimination based on recall; in contrast, list strength had no effect whatsoever on familiarity-based discrimination.[2] We should emphasize that the technique we used to estimate familiarity-based discrimination assumes that recall and familiarity are independent–the independence assumption is discussed in more detail in *Simulation 7: The Combined Model and the Independence Assumption*, below. Also, as discussed by Norman, one cannot conclusively rule out (in this case) the possibility that the observed LSE for recall-based discrimination was caused by shifting response bias, with no real change in sensitivity. Nonetheless, the overall pattern of results is highly consistent with the predictions of our model.

*Lure relatedness.* Norman (2002), also looked at how list strength affected discrimination in the plurals paradigm (Hintzman et al., 1992), where participants have to discriminate between studied words, 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 (see *Simulation 3: YN Related-Lure Simulations*, above). Thus, we should find a significant LSE for studied versus SP discrimination but not necessarily for studied

[2] Norman (2002) did not report how list strength affects familiarity-based discrimination—these results are being presented for the first time here.

versus unrelated discrimination, which can also be supported by familiarity.

Furthermore, we can also look at SP versus unrelated pseudodiscrimination, that is, how much more likely participants are to say *old* to related versus unrelated lures. Familiarity boosts pseudodiscrimination (insofar as SP lures are more familiar than unrelated lures), but recall of plurality information lowers pseudodiscrimination (by allowing participants to confidently reject SP lures). Hence, if increasing list strength reduces recall of plurality information (but has no effect on familiarity-based discrimination), the net effect should be an increase in pseudodiscrimination (a negative LSE).

As predicted, the LSE for studied versus SP-lure discrimination is significant, the LSE for studied versus unrelated-lure discrimination is nonsignificant, and there is a significant negative LSE for SP-lure versus unrelated-lure pseudodiscrimination (see Figure 24).

In summary, we obtained data consistent with the model's prediction of an LSE for recall-based discrimination and with a null LSE for familiarity-based discrimination, using two very different methods of isolating the contributions of these processes (collecting self-report data and using related lures). The model's predictions regarding the boundary conditions of the null LSE for familiarity-based discrimination remain to be tested. For example, our claim that discrimination asymptotically goes to zero with increasing list strength implies that it should be possible to obtain an LSE (in paradigms that have previously yielded a null LSE) by increasing the number of training trials for interference items.

### *List-Length Effects*

Thus far, our interference simulations have focused on list strength. Here, we briefly discuss how list length affects performance in the two models. In contrast to the list-strength paradigm, which measures how strengthening items already on the list interferes with memory for other items, the list-length paradigm measures how adding new (previously nonstudied) items to the list interferes with memory for other items.

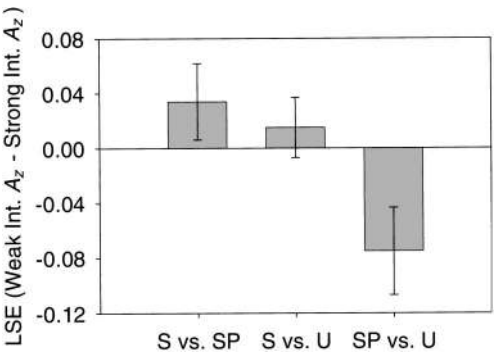


*Figure 24.* Results from the plurals list-strength effect (LSE) experiment. In this experiment, recognition sensitivity was measured using $A_z$ (an index of the area under the receiver-operating characteristic curve; Macmillan & Creelman, 1991). The graph plots the size of the LSE for three different kinds of discrimination: Studied (S) versus related switched-plurality (SP) lures; S versus unrelated (U) lures; and SP versus U lure pseudodiscrimination. Error bars indicate 95% confidence intervals. There was a significant LSE for S versus SP lure discrimination, and there was a significant negative LSE for SP versus U lure pseudodiscrimination. Strong int. = strong interference; Weak int. = weak interference.

The basic finding is that our model, as currently configured, makes the same predictions regarding list-length and list-strength effects—adding new items and strengthening already-presented items induce a comparable amount of weight change and therefore result in comparable levels of interference. As with list strength, the model predicts a dissociation whereby (so long as overlap between items is not too high) list length affects discrimination based on hippocampal recall but not MTLC familiarity. Yonelinas (1994) obtained evidence consistent with this prediction using the process-dissociation procedure (which assumes independence).

However, some extant evidence appears to contradict the model's prediction of parallel effects of list length and list strength. In particular, Murnane and Shiffrin (1991a) and others have obtained a dissociation whereby list length impairs recognition sensitivity but a closely matched list-strength manipulation does not. This finding implies that adding new items to the list is more disruptive to existing weight patterns than repeating already-studied items. We are currently exploring different ways of implementing this dynamic in our model.

One particularly promising approach is to add a large, transient fast-weight component to the model that reaches its maximum value in one study trial and then decays exponentially; subsequent presentations of the same item simply reset fast weights to their maximum value, and decay begins again (see Hinton & Plaut, 1987, for an early implementation of a similar mechanism). This dynamic is consistent with neurobiological evidence showing a large, transient component to LTP (see, e.g., Bliss & Collingridge, 1993; Malenka & Nicoll, 1993).

The assumptions outlined above imply that the magnitude of fast weights at test (for a particular item) is a function of the time elapsed from the most recent presentation of the item; the number of times that the item has been repeated before this final presentation does not matter. As such, presenting interference items for the first time (increasing list length) should have a large effect on fast weights, but repeating already-studied interference items (increasing list strength) should have less of an effect on the configuration of fast weights at test. Preliminary cortical-model simulations incorporating fast weights (in addition to standard, nondecaying weights) have shown a list-length–list-strength dissociation, but more work is needed to explore the relative merits of different implementations of fast weights (e.g., should the fast-weight component incorporate LTD as well as LTP?) and how the presence of fast weights interacts with the other predictions outlined in this article.

The idea that list-length effects are attributable to quickly decaying weights implies that interposing a delay between study and test (thereby allowing the fast weights to decay) should greatly diminish the list-length effect. In keeping with this prediction, a recent study with a 5-min 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 7: The Combined Model and the Independence Assumption*

Up to this point, we have explored the properties of hippocampal recall and MTLC familiarity by presenting input patterns to separate hippocampal and neocortical 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 article, all extant techniques for measuring the distinct contributions of recall and familiarity to recognition performance assume that they are stochastically independent (see, e.g., Jacoby et al., 1997). This assumption cannot be directly tested using behavioral data because of the chicken-and-egg problems described in the introductory section, above.

To assess the validity of the independence assumption, we implemented a combined model in which the cortical system serves as the input to the hippocampus—this arrangement more accurately reflects how the two systems are connected in the brain. Using the combined model, we show here that there is no simple answer regarding whether or not hippocampal recall and MTLC familiarity are independent. Rather, the extent to which these processes are correlated is a function of different (situationally varying) factors, some of which boost the correlation and some of which reduce the correlation. In this section, we briefly describe the architecture of the combined model, and then we use the model to explore two manipulations that push the recall–familiarity correlation in opposite directions: encoding variability and interference (list length).

#### *Combined-Model Architecture*

The combined model is structurally identical to the hippocampal model except that the projection from input to EC_in has modifiable connections (and 25% random connectivity) instead of fixed one-to-one 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 (i.e., the EC_in layer of the combined model serves the same functional role as the MTLC layer of the separate cortical network).

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 has only 240 units, compared with 1,920 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 that we used input patterns with 32.5% overlap—this level of input overlap yields approximately 24% overlap between EC_in patterns at study.

In the combined model, the absolute size of the recall–familiarity correlation is likely to be inflated, relative to the brain, because of the high level of sampling variability present in our model. Sampling variability leads to random fluctuations in the sharpness of cortical representations, which induce a correlation (insofar as sharper representations trigger larger familiarity scores, and they also propagate better into the hippocampus, bolstering recall). Because of this issue, our simulations below focus on identifying manipulations that affect the size of the correlation, rather than characterizing the absolute size of the correlation.

#### *Effects of 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, whereas 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. For each simulated participant, we read out 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) on each trial and measured the correlation between these signals (across trials). The results of these simulations are presented in Figure 25; as expected, increasing encoding variability increased the recall–familiarity correlation in the model.

#### *Interference-Induced Decorrelation*

In the next set of simulations, we show how the presence of interference between memory traces can reduce the recall–familiarity correlation. In *Simulation 6: Interference and List Strength*, above, we discussed how the two systems are

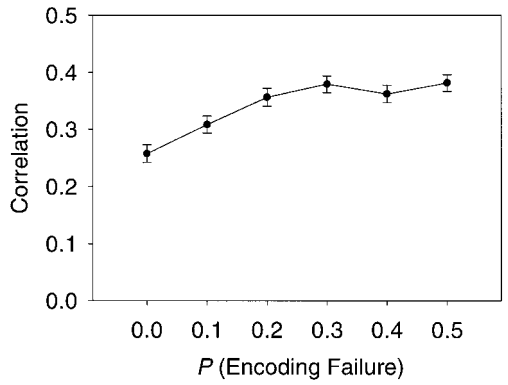


*Figure 25.* Simulations exploring the effect of encoding variability on the recall–familiarity correlation for studied items. As encoding variability (operationalized as the probability that the model will "blink" and fail to encode half of an item's features at study) increases, the recall–familiarity correlation increases.

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-items 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). There were 10 target items followed by 0 to 150 (nontested) interference items. As expected, increasing list length lowers the recall–familiarity correlation for studied items (see Figure 26A) in the model.

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 (see Figure 26B). With increasing interference, recall decreases sharply, but familiarity stays relatively constant; this differential effect of interference works to decorrelate the two signals. 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.

In summary, the combined model gives us a principled means of predicting how different factors affect the statistical relationship between recall and familiarity. Given the tight coupling of the cortical and hippocampal networks in the combined model, one might think that a positive correlation is inevitable. However, the results presented here show that—to a first approximation—independence can be achieved so long as factors that reduce the correlation (e.g., interference) exert more of an influence than factors that bolster the correlation (e.g., encoding variability). Future research will focus on identifying additional factors that affect the size of the correlation.

### *Simulation 8: 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. Specifically, we show that (in the combined model) partial hippocampal lesions can sometimes result in worse overall recognition performance than complete hippocampal lesions. In contrast, increasing MTLC lesion size monotonically reduces overall recognition performance.

### *Partial Lesion Simulation*

*Method.* We ran one set of simulations exploring the effects of focal hippocampal lesions and another set of simulations exploring the effects of focal MTLC lesions. In all of the lesion simulations, the size of the lesion (in terms of percentage 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 (prelesion) 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 these simulations, we used FC testing to maximize comparability with animal lesion studies that had used this format (see, e.g., Baxter & Murray, 2001b).

To simulate overall recognition performance when both processes are contributing, we used a decision rule whereby if one item triggered a larger positive recall score (*match* − *mismatch*) than the other, then that item was selected; otherwise, if *match* − *mismatch* is less than or equal to 0, the decision fell back on familiarity. This rule incorporates the assumption (shared by other dual-process models; e.g., Jacoby et al., 1997) that recall takes precedence over familiarity. This differential weighting of recall can be justified in terms of our finding that, in normal circumstances, hippocampal recall in the CLS model is more diagnostic than MTLC familiarity. Furthermore, it is supported by data showing that recall is associated with higher average recognition confidence ratings than familiarity (see, e.g., Yonelinas, 2001).

*Hippocampal lesion results.* Figure 27 shows how hippocampal FC performance, cortical FC performance, and combined FC

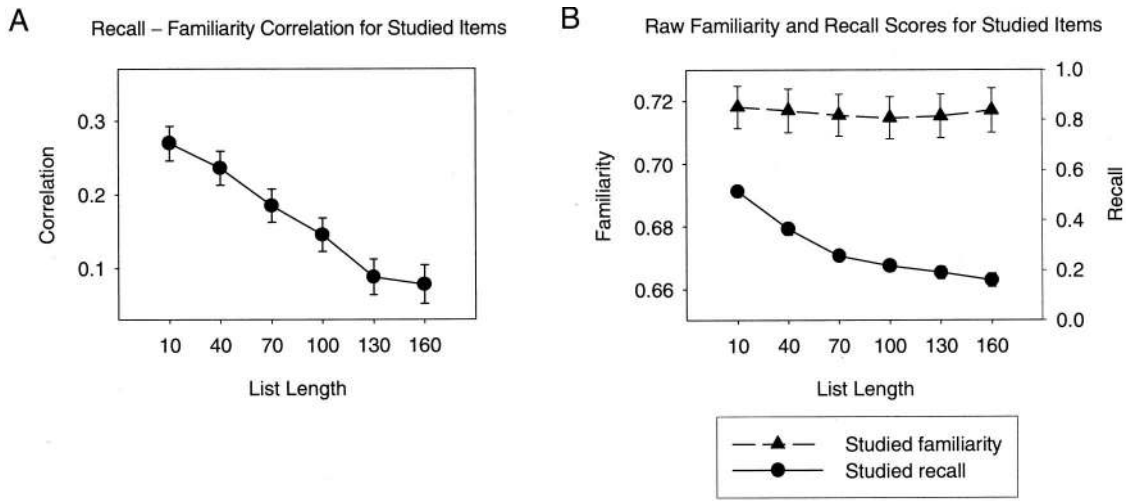


*Figure 26.* Simulations exploring how interference affects the recall–familiarity correlation. A: Increasing list length reduces the recall–familiarity correlation for studied items. B: Increasing list length leads to a decrease in studied-item recall scores, but studied-item familiarity scores stay relatively constant.

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, whereas cortical performance is unaffected by hippocampal damage (because familiarity is computed before activity is fed into the hippocampus). The most interesting result is that hippocampal lesion size has a nonmonotonic effect on combined recognition performance. At first, combined FC accuracy decreases with increasing hippocampal lesion size; however, going from a 75% to a 100% hippocampal lesion actually improves combined FC performance.

Why is recognition performance worse for partial (75%) versus complete lesions? The key to understanding this finding is that partial hippocampal damage impairs the hippocampus's 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; accordingly, 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.

Pattern-separation failure (induced by partial damage) results in a sharp increase in the amount of recall triggered by lures. Combined recognition performance in these participants suffers because the noisy recall signal drowns out useful information that is present in the familiarity signal. Moving from a partial hippocampal lesion to a complete lesion improves performance by removing this source of noise.

Figure 28 provides direct support for the noisy-hippocampus theory; it shows how the probability that lures will trigger a positive *match − mismatch* score increases steadily with increasing hippocampal lesion size until it reaches a peak of .41 (for 55% hippocampal damage). However, as lesion size approaches 60%, the probability that lures and studied items will trigger a positive *match − mismatch* score starts to decrease. This occurs for two reasons: First, because of pattern-separation failure in CA3, all items start to trigger recall of the same prototypical features (which mismatch item-specific features of the test probe); second, CA1 damage reduces the hippocampus's ability to translate CA3 activity into the target EC pattern. As the number of items triggering positive match–mismatch scores decreases, 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.

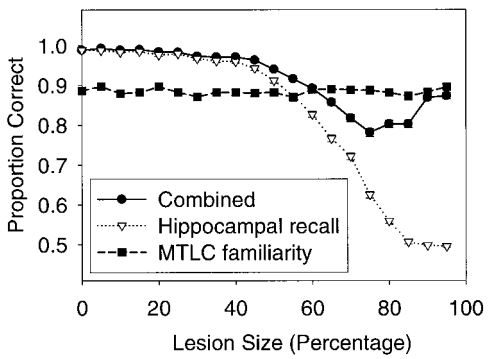


*Figure 27.* Effect of hippocampal damage on forced-choice (FC) recognition performance. This graph plots FC accuracy based on medial temporal lobe cortex (MTLC) familiarity, hippocampal recall, and a combination of the two signals, as a function of hippocampal lesion size. FC accuracy based on MTLC familiarity is unaffected by hippocampal lesion size. 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: Initially, combined FC accuracy declines; however, going from a 75% to a 100% hippocampal lesion leads to an increase in combined FC performance.

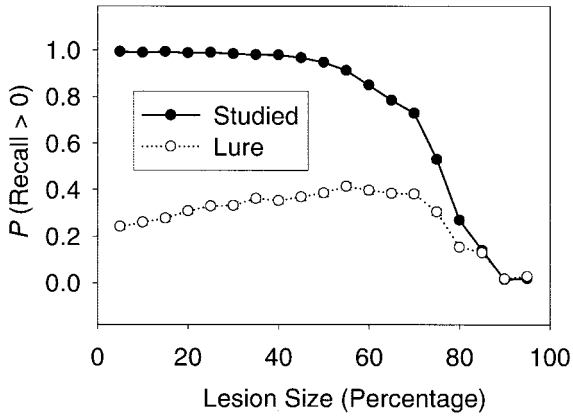


*Figure 28.* Plot of the probability that studied items and lures will trigger a positive *match − mismatch* score, as a function of hippocampal lesion size. For studied items, the probability declines monotonically as a function of lesion size. For lures, the probability first increases, then decreases, as a function of lesion size.

*MTLC lesion results.* Turning to the effects of MTLC lesions, we found that lesioning EC_in hurts both cortical and hippocampal recognition performance (see Figure 29). Therefore, combined recognition performance decreases steadily as a function of MTLC lesion size. 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.

### *Relevant Data and Implications*

The simulation results reported above are consistent with the results of a recent meta-analysis conducted by Baxter and Murray (2001b). This meta-analysis incorporated results from several studies that have looked at hippocampal and perirhinal (MTLC) lesion effects on recognition in monkeys using a delayed-nonmatching-to-sample paradigm. Baxter and Murray found, in keeping with our results, that partial hippocampal lesions can lead to larger recognition deficits than more complete lesions—in the meta-analysis, hippocampal lesion size and recognition impairment were negatively correlated. The Baxter and Murray meta-analysis also found that perirhinal lesion size and recognition impairment were

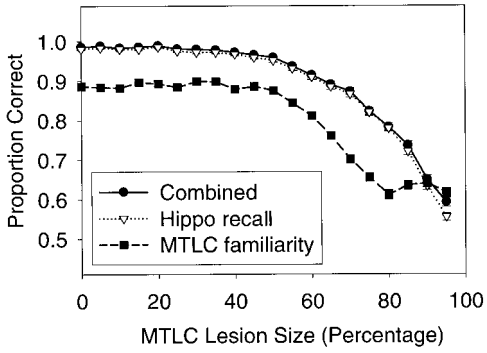


*Figure 29.* Effect of medial temporal lobe cortex (MTLC)—specifically, entorhinal cortex input layer (EC_in)—damage on forced-choice (FC) recognition performance. This graph plots FC accuracy based on MTLC familiarity, hippocampal recall, and a combination of the two signals, as a function of EC_in lesion size. All three accuracy scores (recall alone, familiarity alone, and combined) decline steadily with increasing lesion size. Hippo = hippocampus.

positively correlated, just as we found that MTLC lesion size and impairment were correlated in our simulations.

Baxter and Murray's (2001b) results are highly controversial; Zola and Squire (2001) reanalyzed the data in the Baxter and Murray meta-analysis using a different set of statistical techniques that control, for example, for differences in mean lesion size across studies and found that the negative correlation between hippocampal lesion size and impairment reported by Baxter and Murray was no longer significant (although there was still a nonsignificant trend in this direction; see Baxter & Murray, 2001a, for further discussion of this issue). Our results contribute to this debate by providing a novel and principled account of how a negative correlation might come into being in terms of pattern-separation failure resulting in a noisy recall signal that participants nevertheless relied on when making recognition judgments. Of course, our account is not the only way of explaining why partial hippocampal lesions might impair recognition more than complete lesions. Another possibility, suggested by Mumby et al. (1996), is that a damaged hippocampus might disrupt neocortical processing via epileptiform activity.

The results reported here speak to the debate over why hippocampally lesioned patients sometimes show relatively spared recognition performance on standard recognition tests with unrelated lures and sometimes do not. The model predicts that patients with relatively complete hippocampal lesions (that nonetheless spare MTLC) should show relatively spared performance and patients with smaller hippocampal lesions (that reduce the diagnosticity of recall without eliminating the signal outright) should show an across-the-board declarative memory deficit without any particular sparing of recognition. This view implies that the range of etiologies that produce selective sparing should be quite narrow: If the lesion is too small, one ends up with a partially lesioned hippocampus that injects noise into the recognition process; if the lesion is too large, then one hits perirhinal cortex (in addition to the hippocampus), and this leads to deficits in familiarity-based recognition.

## General Discussion

In this section of the article, we review how our modeling work speaks to extant debates over the characterization of the respective contributions of recall (vs. familiarity) and hippocampus (vs. MTLC) to recognition. Next, we compare our model to other neurally inspired and abstract computational models of recognition. We conclude by discussing limitations of the models and future directions for research.

### *Implications for Theories of How Recall Contributes to Recognition*

As discussed in the introductory section, above, the dual-process approach to recognition has become increasingly prevalent in recent years, but this enterprise is based on a weak foundation. Yonelinas, Jacoby, and their colleagues (see Yonelinas, 2001) have developed several techniques for measuring the contributions of recall and familiarity based on behavioral data, but these techniques rely on a core set of assumptions that have not been tested; furthermore, some of these assumptions (e.g., independence) cannot be tested based on behavioral data alone because of chicken-and-egg problems.

More specifically, measurement techniques such as process dissociation and ROC analysis are built around dual-process signal-detection theory (Jacoby et al., 1997; Yonelinas, 2001). In addition to the independence assumption, this theory assumes that familiarity is a Gaussian signal-detection process but recall is a dual high-threshold process (i.e., recall is all or none; studied items are sometimes recalled as being *old,* but lures never are; lures are sometimes recalled as being *new,* but studied items never are). The CLS model does not assume any of the above claims to be true—rather, its core assumptions are based on the functional properties of hippocampus versus MTLC. As such, we can use the CLS model to evaluate the validity of dual-process signal-detection theory.

Results from our cortical model are generally consistent with the idea that familiarity is a Gaussian signal-detection process. In contrast, results from our hippocampal model are not consistent with the idea that recall is a high-threshold process—recall in our model is not all or none, and lures sometimes trigger above-zero recall. Our claims that lures can trigger some matching recall and that the number of recalled details varies from item to item are consistent with the source monitoring framework set forth by Marcia Johnson and her colleagues (see, e.g., Mitchell & Johnson, 2000).

Although the recall process in our model is not strictly high threshold, it is not Gaussian either. If overlap between stimuli is not too high, our recall process is approximately consistent with dual high-threshold theory in the sense that (for a given experiment) there is a level of matching recall that is sometimes exceeded by studied items but not by lures and there is a level of mismatching recall that is sometimes exceeded by lures but not by studied items. Furthermore, we assume that participants routinely set their recall criterion high enough to avoid false recognition of unrelated lures and that participants set their criterion for recall-to-reject high enough to avoid incorrect rejection of studied items. As such, the model provides some support (conditional on overlap not being too high) for Yonelinas's (see Yonelinas et al., 1996)

claim that lures will not be called *old* based on recall and that studied items will not be called *new* based on recall.

Regarding the independence assumption, as mentioned earlier, Curran and Hintzman (1995) and others have criticized this assumption on the grounds that some degree of encoding variability is likely to be present in any experiment, and encoding variability results in a positive recall–familiarity correlation. Our results confirm this latter conclusion, but they also show that interference reduces the recall–familiarity correlation. As such, it is possible to achieve independence (assuming there is enough interference) even when encoding variability is present.

Overall, although the details of the CLS model's predictions are not strictly consistent with the assumptions behind dual-process signal-detection theory, it is safe to say that the CLS model's predictions are broadly consistent with these assumptions; accordingly, we would expect measurement techniques that rely on these assumptions to yield meaningful results most of the time. The main contribution of the CLS model is to establish boundary conditions on the validity of these assumptions—for example, the assumption that lures will not be called *old* based on recall does not hold true when there is extensive overlap between stimuli, and the independence assumption is more likely to hold true in situations where interference is high and encoding variability is low than when the opposite is true.

### *Implications for Theories of How the Hippocampus (Versus the MTLC) Contributes to Recognition*

To a first approximation, the CLS model resembles the neuropsychological theory of episodic memory set forth by Aggleton and Brown (1999; hereafter, the A&B theory)—both theories posit that the hippocampus is essential for recall and that MTLC can support judgments of stimulus familiarity on its own. However, this resemblance is only superficial. As discussed in the introductory section, above, simply linking *recall* and *familiarity* to hippocampus and MTLC, without further specifying how these processes work, does not allow one to predict when recognition will be impaired or spared after hippocampal damage. Everything depends on how one unpacks the terms *recall* and *familiarity,* and the CLS and A&B theories unpack these terms in completely different ways.

The A&B theory unpacks recall and familiarity in terms of a simple, verbally stated dichotomy whereby recall is necessary for forming new associations but familiarity is sufficient for item recognition. In contrast to the A&B theory, the CLS model grounds its conception of recall and familiarity in terms of very specific ideas about the neurocomputational properties of hippocampus and MTLC. Moreover, we have implemented these ideas in a working computational model that can be used to test their sufficiency and to generate novel predictions.

According to the CLS model, practically all differences between cortical and hippocampal contributions to recognition can be traced back to graded differences in how information is represented in these structures. For example, the fact that hippocampal representations are more sparse than cortical representations in our model implies that hippocampal recall is more sensitive to conjunctions than MTLC familiarity, but crucially, both signals should show some sensitivity to conjunctions (see *Simulation 5: Associative Recognition and Sensitivity to Conjunctions*, above, for more discussion of this point). This graded approach makes it possible for the CLS model to explain dissociations within simple categories such as memory for new associations—for example, the finding from Mayes et al. (2001) that hippocampally lesioned patient Y.R. sometimes showed intact FC word–word associative recognition (presumably, based on MTLC familiarity) despite being impaired at cued recall of newly learned paired associates (Mayes et al., 2002; see Vargha-Khadem et al., 1997, for a similar pattern of results). As discussed earlier, these dissociations are problematic for the A&B theory's item vs. associative dichotomy.

Lastly, we should mention an important commonality between the CLS model and the A&B theory: Both theories predict sparing of item recognition (with unrelated lures) relative to recall in patients with complete hippocampal lesions. Although some studies have found this pattern of results in patients with focal hippocampal damage (see, e.g., Mayes et al., 2002), it is potentially quite problematic for both models that other studies have found roughly equivalent deficits in recognition (with unrelated lures) and recall following focal hippocampal damage (see, e.g., Manns & Squire, 1999). We do not claim to fully understand why some hippocampally lesioned patients have shown relatively spared item recognition but others have not. 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 *Simulation 8: Lesion Effects in the Combined Model*, above, that partial hippocampal lesions are especially harmful to recognition. This idea is still highly controversial and needs to be tested directly, using experiments that carry out careful, parametric manipulations of lesion size (controlling for other factors such as lesion technique and task).

### *Comparison With Abstract Computational Models of Recognition*

Whereas our model incorporates explicit claims about how different brain structures (hippocampus and MTLC) support recognition memory, most computational models of recognition memory are abstract in the sense that they do not make specific claims about how recognition is implemented in the brain. The REM (i.e., retrieving effectively from memory) model presented by Shiffrin and Steyvers (1997) represents the state of the art in abstract modeling of recognition memory (see McClelland & Chappell, 1998, for a very similar model). REM carries 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 REM and other 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 such as REM posit that differentiation (Shiffrin et al., 1990) contributes to the null LSE. Thus, our modeling work relies critically on insights that were developed in the context of abstract models such as REM.

We can also draw a number of contrasts between the CLS model and abstract Bayesian models. One major difference is that our model posits that two processes (with distinct operating characteristics) contribute to recognition whereas abstract models attempt to explain recognition data in terms of a single familiarity process.

Another difference is that—in our model—interference occurs at study, when one item reuses weights that are also used by another item, whereas REM posits that memory traces are stored in a noninterfering fashion and that interference arises at test, whenever the test item spuriously matches memory traces corresponding to other items.

The question of whether interference occurs at study (or only at test) has a long history in memory research. Other models that posit structural interference at study have found large and sometimes excessive effects of interference on recognition sensitivity. For example, Ratcliff (1990) presented a model consisting of a three-layer feedforward network that learns to reproduce input patterns in the output layer; the dependent measure at test is how well the recalled (output) pattern matches the input. Like our hippocampal model, the Ratcliff model shows interference effects on recognition sensitivity because recall of discriminative features of lures is at floor; as such, any decrease in studied recall necessarily pushes the studied and lure recall distributions closer together. The Ratcliff model generally shows more interference than our hippocampal model because there is more overlap between hidden representations in the Ratcliff model than in our hippocampal model.

On the basis of these results (and others like them), Murnane and Shiffrin (1991a) concluded that interference-at-study models may be incapable of explaining the null recognition LSE obtained by Ratcliff et al. (1990). An important implication of the work presented here is that interference-at-study models do not always show excessive effects of interference on recognition sensitivity; our cortical model predicts a null recognition LSE because increasing list strength reduces lure familiarity slightly more than studied-item familiarity, so the studied–lure gap in familiarity is preserved.

In this article, 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 versus models such as REM. 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 can 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.

## *Comparison With Other Neural Network Models of Hippocampal and Cortical Contributions to Recognition*

### *Models of the Hippocampus*

The hippocampal component of the CLS model is part of a long tradition of hippocampal modeling (see, e.g., Burgess & O'Keefe, 1996; Hasselmo & Wyble, 1997; Levy, 1989; Marr, 1971; McNaughton & Morris, 1987; Moll & Miikkulainen, 1997; Rolls, 1989; Touretzky & Redish, 1996; Treves & Rolls, 1994; Wu, Baxter, & Levy, 1996). 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 have built 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, the H&W model) 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 that the H&W model makes a distinction between item and (shared) context information and posits 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 context 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 LSE 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 effects and LSEs should be obtained in the hippocampus). Thus, because the H&W model uses a different hippocampal recognition measure, as well as separate item and context representations, it generates recognition predictions that are very different from the CLS hippocampal model's predictions. However, we should emphasize that, if the H&W model used the same recognition measure as our model (*match* − *mismatch*) and factored item recall as well as context recall into recognition decisions, it and the CLS model would likely make very similar predictions because the two model architectures are so similar.

### *Models of Neocortical Contributions to Recognition Memory*

In recent years, several models besides ours have been developed that address the role of MTLC in familiarity discrimination; see Bogacz and Brown (2003) for a detailed comparison of the properties of different familiarity-discrimination models. Some of 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 (see, e.g., Bogacz, Brown, & Giraud-Carrier, 2001; Sohal & Hasselmo, 2000), although the details of these models differ from ours (e.g., the Bogacz et al.,

2001, and Sohal & Hasselmo, 2000, models posit that both homosynaptic Hebbian LTD—which decreases weights if the sending unit is active but the receiving unit is not—and heterosynaptic Hebbian LTD—which decreases weights if the receiving unit is active but the sending unit is not—are important for familiarity discrimination, whereas our model incorporates only heterosynaptic Hebbian LTD). Other models incorporate radically different mechanisms, for instance, anti-Hebbian learning that reduces connection strengths between coactive neurons (Bogacz & Brown, 2003).

One difference between the models proposed by Bogacz and Brown (2003) and our model is that—in the Bogacz and Brown models—familiarity is computed by a specialized population of novelty-detector units that are not directly involved in extracting and representing stimulus features (see Bogacz & Brown, 2003, for a review of empirical evidence that supports this claim). In contrast, our combined model does not posit the existence of specialized novelty-detector units; rather, the layer (EC_in) where the *act_win* familiarity signal is read out contains the highest (most refined) cortical representation of the stimulus, which in turn serves as the input to the hippocampus.

We should note, however, that our model could easily be transformed into a model with specialized novelty detectors without altering any of the predictions outlined in the main body of this article. To do this, we could connect the input layer of the combined model directly to CA3, CA1, and DG, and we could have this input layer feed in parallel to a second cortical layer (with no connections to the hippocampus) where *act_win* is computed. The units in this second cortical layer could be labeled specialized novelty detectors insofar as they are not serving any other important function in the model. This change would not affect the functioning of the cortical part of the model in any way.

Having the same layer serve as the input to the novelty-detection layer and the hippocampus could, in principle, affect the predictions of the combined model, but as a practical matter, none of the combined model predictions outlined in Simulations 7 and 8, above, would be affected by this change. For example, if cortical units involved in computing novelty/familiarity were not involved in passing features to the hippocampus, then it would be possible in principle to disrupt familiarity for a particular stimulus without disrupting hippocampal recall of that stimulus (by lesioning the novelty-detector units). However, this is probably not possible in practice—according to Brown and Xiang (1998), perirhinal neurons involved in familiarity discrimination and stimulus representation are topographically interspersed, so lesions large enough to affect one population of neurons should also affect the other.

A major issue raised by Bogacz and Brown (2003) is whether networks such as ours that extract features via Hebbian learning have adequate capacity to explain people's ability to discriminate between very large numbers of familiar and unfamiliar stimuli (e.g., Standing, 1973, found that people can discriminate between studied and nonstudied pictures after studying a list of thousands of pictures). Bogacz and Brown argued that—even in a "brain-sized" version of our cortical model—the network's tendency to represent shared (prototypical) features at the expense of features that discriminate between items will result in unacceptably poor performance after studying large numbers of stimuli. Bogacz and Brown pointed out that the anti-Hebbian model that they proposed does not have this problem; this model ignores features that are shared across patterns and, thus, has a much higher capacity. A possible problem with the anti-Hebbian model is that it may show too little interference. More research is needed to assess whether our network architecture, suitably scaled, can explain findings like those of Standing (1973) and—if not—how it can be modified to accommodate this result (without eliminating its ability to explain interference effects on recognition discrimination).

At this point in time, it is difficult to directly compare our model's predictions with the predictions of other cortical familiarity models because the models have been applied to different data domains—we have focused on explaining detailed patterns of behavioral data, whereas the other models have focused on explaining single-cell recording data in monkeys. Bringing the different models to bear on the same data points is an important topic for future research. Although the CLS model cannot make detailed predictions about spiking patterns of single neurons, it does make predictions regarding how firing rates will change as a function of familiarity. For example, the model predicts that, for a particular stimulus, neurons that show decreased (vs. asymptotically strong) firing in response to repeated presentation of that stimulus should be neurons that initially had a less strong response to the stimulus (and therefore lost the competition to represent the stimulus).

### *Future Directions*

Future research will address limitations of the model that were mentioned earlier. In the Sources of Variability section, above, we discussed how the model incorporates some sources of variability that we plan to remove (sampling variability) and lacks some sources of variability that we plan to add. Increases in computer processing speed will make it possible to expand our networks to the point where sampling variability is negligible, and we will replace lost sampling variability by adding encoding variability and variability in preexperimental presentation frequency to the model. Including preexperimental 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 stimuli than for high-frequency stimuli but false alarms tend to be higher for high-frequency stimuli than for low-frequency stimuli (see, e.g., Glanzer, Adams, Iverson, & Kim, 1993); recently, several studies have obtained evidence suggesting that recall is responsible for the low-frequency hit-rate advantage and familiarity is responsible for the high-frequency false-alarm-rate advantage (Joordens & Hockley, 2000; Reder et al., 2000; Reder et al. also presented an abstract dual-process model of this finding).

Furthermore, we plan to directly address the question of how participants make decisions based on recall and familiarity. Clearly, people are capable of using 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 participants' actual performance. A specific idea that we plan to explore in detail is that participants 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 (B.G.) 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 B.G. 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 participants 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 (see, e.g., Jacoby, 1991). We could also begin to address the rich literature on how different manipulations affect recognition ROC curves (see, e.g., Ratcliff et al., 1992; Yonelinas, 1994).

Another topic for future research involves improving cross talk between the model and neuroimaging data. In principle, we should be able to predict functional magnetic resonance imaging (fMRI) activations during episodic recognition tasks by reading out activation from different subregions of the model; to achieve this goal, we 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 event-related potential (ERP) correlates of recall and familiarity; 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 discriminated between targets and lures, but list strength adversely affected how well the ERP recall correlate discriminated between targets and lures (Norman, Curran, & Tepe, 2002).

We also plan to explore other, more biologically plausible ways of reading out familiarity from the cortical network. Although *act_win* has the virtue of being easy to compute, 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). One promising alternative measure is *settle_time*: the time it takes for activity to spread through the network (more concretely, we measured the number of processing cycles needed for average activity in MTLC to reach a criterion value of .03). This measure exploits the fact that activity spreads more quickly for familiar than for unfamiliar patterns. The *settle_time* measure is more biologically plausible than *act_win* insofar as it requires only 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 predetermined criterion. Preliminary simulation results have shown that *settle_time* yields good $d'$ scores and—like *act_win*—does not show an LSE on $d'$ for our basic parameters (20% overlap). Further research is 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.

## 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 on this foundation to address a range of other human and animal memory phenomena.

## 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* (pp. 121–268). New York: Wiley.

Bauml, 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, January 7). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature, 361,* 31–39.

Bogacz, R., & Brown, M. W. (2003). Comparison of computational models of familiarity discrimination in perirhinal cortex. *Hippocampus, 13,* 494–524.

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience, 10,* 5–23.

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 & 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, England: Psychology Press.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & 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.

Douglas, R. J., & Martin, K. A. C. (1990). Neocortex. In G. M. Shepherd (Ed.), *The synaptic organization of the brain* (pp. 389–438). Oxford, England: 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,* 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.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121–134.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Scwab & H. C. Nusbaum (Eds.), *Pattern recognition in humans and machines: Vol. 1: Speech perception* (pp. 187–294). New York: Academic Press.

Grossberg, S., & Stone, G. (1986). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review, 93,* 46–74.

Hasselmo, M. E. (1995). Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioural Brain Research, 67,* 1–27.

Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation, 14,* 793–818.

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. In *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 & Cognition, 29,* 547–556.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similiarity 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. (2002). Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus, 12,* 341–351.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96,* 208–233.

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). Mahwah, NJ: Erlbaum.

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., Rizzuto, D., & Schneider, A. (2003). *Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models.* Manuscript submitted for publication.

Kanerva, P. (1988). *Sparse distributed memory.* Cambridge, MA: MIT Press.

Kohonen, T. (1977). *Associative memory: A system theoretical approach.* Berlin, Germany: Springer-Verlag.

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 of London, Series B, 262,* 23–81.

Mayes, A. R., Holdstock, J. S., Isaac, C. L., Hunkin, N. M., & Roberts, N. (2002). Relative sparing of item recognition memory in a patient with adult-onset damage limited to the hippocampus. *Hippocampus, 12,* 325–340.

Mayes, A. R., Isaac, C. L., Downes, J. J., Holdstock, J. S., Hunkin, N. M., Montaldi, D., et al. (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., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus, 6,* 654–665.

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.

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences, 10,* 408–415.

Miller, E. K., Li, L., & Desimone, R. (1991, November 29). A neural mechanism for working and recognition memory in inferior temporal cortex. *Science, 254,* 1377–1379.

Minai, A. A., & Levy, W. B. (1994). Setting the activity level in sparse random networks. *Neural Computation, 6,* 85–99.

Mishkin, M., Suzuki, W., Gadian, D. G., & Vargha-Khadem, F. (1997). Hierarchical organization of cognitive memory. *Philosophical Transactions of the Royal Society of London, Series 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.

Mitchell, K. J., & Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 179–195). New York: Oxford University Press.

Moll, M., & Miikkulainen, R. (1997). Convergence-zone episodic memory: Analysis and simulations. *Neural Networks, 10,* 1017–1036.

Mumby, D. G., Wood, E. R., Duva, C. A., Kornecook, T. J., Pinel, J. P. J., & Phillips, A. G. (1996). Ischemia-induced object recognition deficits in rats are attenuated by hippocampal lesions before or soon after ischemia. *Behavioral Neuroscience, 110,* 266–281.

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 & 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.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1083–1094.

Norman, K., Curran, T., & Tepe, K. (2002, April). *Event-related potential correlates of interference effects on recognition memory.* Poster presented at the 9th Annual Meeting of the Cognitive Neuroscience Society, San Francisco.

Nowlan, S. J. (1990). Maximum likelihood competitive learning. In D. S. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2, pp. 574–582). San Mateo, CA: Morgan Kaufmann.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology, 15,* 267–273.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map.* Oxford, England: Oxford University Press.

O'Reilly, R. C. (1996). *The Leabra model of neural interactions and learning in the neocortex.* Unpublished doctoral dissertation, Carnegie Mellon University.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences, 2,* 455–462.

O'Reilly, R. C., & Busby, R. S. (2002). Generalizable relational binding from coarse-coded distributed representations. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (Vol. 14, pp. 75–82). Cambridge, MA: MIT Press.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus, 4,* 661–682.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by stimulating 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* (Vol. 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,* 1763–1779.

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., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject. *Journal of Memory and Language, 40,* 432–453.

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28,* 907–922.

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. J. (1989). The hippocampal formation is necessary for rats to learn and remember configural discriminations. *Behavioural Brain Research, 34,* 97–109.

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: Vol. 1. Foundations* (pp. 151–193). Cambridge, MA: MIT Press.

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review, 94,* 439–454.

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., Ratcliff, R., & Clark, S. E. (1990). The list-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179–195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory:

REM—Retrieving effectively from memory. *Psychonomic Bulletin & 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,* 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. (1992a). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience, 4,* 232–243.

Squire, L. R. (1992b). 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-Morgan, S. (1991, September 20). The medial temporal memory system. *Science, 253,* 1380–1386.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25,* 207–222.

Stark, C. E. L., & Squire, L. R. (2002, April). *Memory for associations in patients with hippocampal damage.* Poster session presented at the 9th Annual Meeting of the Cognitive Neuroscience Society, San Francisco.

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

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.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997, July 18). 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. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25,* 747–763.

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General, 130,* 361–379.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517.

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & 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.

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., Buffalo, 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.

(*Appendixes follow*)

# Appendix A

# Algorithm Details

This appendix describes the computational details of the algorithm that was used in the simulations. The algorithm is identical to the Leabra algorithm described in O'Reilly and Munakata (2000; see also O'Reilly, 1998), except that the error-driven learning component of the Leabra algorithm was not used here; see Grossberg (1976) for a similar model. Interested readers should refer to O'Reilly and Munakata for more details regarding the algorithm and its historical precedents.

## Pseudocode

The pseudocode for the algorithm that we used 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, settle over cycles of updating:

1. At start of settling, for all units:
   A. Initialize all state variables (activation, *v_m*, etc.).
   B. Apply external patterns.
2. During each cycle of settling, for all nonclamped units:
   A. Compute excitatory netinput ($g_e(t)$ or $\eta_j$, Equation A3).
   B. Compute *k*WTA inhibition for each layer based on $g_i^\Theta$ (Equation A6):
      i. Sort units into two groups based on $g_i^\Theta$: top *k* and remaining $k + 1$ to *n*.
      ii. Set inhibitory conductance $g_i$ between $g_k^\Theta$ and $g_{k+1}^\Theta$ (Equation A5).
   C. Compute point-neuron activation combining excitatory input and inhibition (equation A1).
3. Update the weights (based on linear current weight values) for all connections:
   A. Compute Hebbian weight changes (Equation A7).
   B. Increment the weights and apply contrast-enhancement (Equation A9).

## 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. Furthermore, 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:

$$\frac{dV_m(t)}{dt} = \tau \sum_c g_c(t)\bar{g}_c(E_c - V_m(t)) , \tag{A1}$$

with three channels ($c$) corresponding to excitatory input ($e$), leak current ($l$), and inhibitory input ($i$). 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 $E_i$) to 0:

$$V_m^\infty = \frac{g_e\bar{g}_e}{g_e\bar{g}_e + g_l\bar{g}_l + g_i\bar{g}_i} , \tag{A2}$$

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:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij}. \tag{A3}$$

The inhibitory conductance is computed via the *k*-winners-take-all (*k*WTA) 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 $\chi$:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma[V_m(t) - \Theta]_+}\right)} , \tag{A4}$$

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*WTA Inhibition

Leabra uses a *k*WTA function to achieve sparse distributed representations (cf., Minai & Levy, 1994). Although two different versions are possible (see O'Reilly & Munakata, 2000, for details), only the simpler form was used in the present simulations. A uniform level of inhibitory current for all units in the layer is computed as follows:

$$g_i = g_{k+1}^\Theta + q(g_k^\Theta - g_{k+1}^\Theta), \tag{A5}$$

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:

$$g_i^\Theta = \frac{g_e^*\bar{g}_e(E_e - \Theta) + g_l\bar{g}_l(E_l - \Theta)}{\Theta - E_i} , \tag{A6}$$

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

In the basic version of the *k*WTA function used here, $g_k^\Theta$ 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. We should emphasize that, when the membrane potential is at threshold, unit activation in the model = .25; as such, the *k*WTA algorithm places a firm upper bound on the number of units showing activation greater than .25, but it does not set an upper bound on the number of weakly active units (i.e., units showing activation between 0 and .25).

Activation dynamics similar to those produced by the *k*WTA function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition (O'Reilly & Munakata, 2000). Thus, although the *k*WTA 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 (Grossberg, 1976; Nowlan, 1990; Rumelhart & Zipser, 1986), which can be seen as a variant of the Oja normalization (Oja, 1982):

$$\Delta_{hebb} w_{ij} = x_i y_j - y_j w_{ij} = y_j (x_i - w_{ij}). \quad \text{(A7)}$$

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 A7 can be rewritten as follows:

$$\Delta w_{ij} = \varepsilon [y_j x_i (m - w_{ij}) + y_j (1 - x_i)(0 - w_{ij})], \quad \text{(A8)}$$

where an $m$ value of 1 gives Equation A7, whereas a larger value can ensure that the weight value between uncorrelated but sparsely active units is around .5. Specifically, we set $m = .5/\alpha_m$ 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 introduced 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:

$$\hat{w}_{ij} = \frac{1}{1 + \left( \frac{w_{ij}}{\theta (1 - w_{ij})} \right)^{-\gamma}}, \quad \text{(A9)}$$

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

## Hippocampal Model Details

This section provides a brief summary of key architectural parameters of the hippocampal model. Activity levels, layer size, and projection parameters were set to mirror the consensus view of the functional architecture of the hippocampus described, for example, by Squire, Shimamura, and Amaral (1989).

Table B1 shows the sizes of different hippocampal subregions and their activity levels in the model. These activity levels are enforced by setting appropriate $k$ parameters in the Leabra $k$-winners-take-all (kWTA) inhibition function. As discussed in the main text, activity is much more sparse in dentate gyrus and Region CA3 than in entorhinal cortex (EC).

Table B2 shows the properties of the four modifiable projections in the hippocampal model. For each simulated participant, connection weights in these projections were set to values randomly sampled from a uniform distribution with mean and variance (range) as specified in the table. The Scale factor listed in the table shows how influential this projection is, relative to other projections coming into the layer, and % Con (*percentage connectivity*) specifies the percentage of units in the sending layer connected to each unit in the receiving layer. Relative to the perforant path, the mossy fiber pathway is sparse (i.e., each CA3 neuron receives a much smaller number of mossy fiber synapses than perforant path synapses) and strong (i.e., a given mossy fiber synapse has a much larger impact on CA3 unit activation than a given perforant path synapse). The CA3 recurrents and the Schaffer collaterals projecting from CA3 to CA1 are relatively diffuse, so that each CA3 neuron and each CA1 neuron receive a large number of inputs sampled from the entire CA3 population.

Table B1
*Sizes of Different Subregions and Their Activity Levels in the Model*

| Area | Units | Activity (%) |
|---|---|---|
| EC | 240 | 10.0 |
| DG | 1,600 | 1.0 |
| CA3 | 480 | 4.0 |
| CA1 | 640 | 10.0 |

*Note.* EC = entorhinal cortex; DG = dentate gyrus; CA3 and CA1 = regions of the hippocampus.

Table B2
*Properties of Modifiable Projections in the Hippocampal Model*

| Projection | Mean | Var | Scale | % Con |
|---|---|---|---|---|
| EC to DG, CA3 (perforant path) | .5 | .25 | 1 | 25 |
| DG to CA3 (mossy fiber) | .9 | .01 | 25 | 4 |
| CA3 recurrent | .5 | .25 | 1 | 100 |
| CA3 to CA1 (Schaffer) | .5 | .25 | 1 | 100 |

*Note.* Mean = mean initial weight strength; Var = variance of the initial weight distribution; Scale = scaling of this projection relative to other projections; % Con = percentage connectivity; EC = entorhinal cortex; DG = dentate gyrus; CA3 and CA1 = regions of the hippocampus.

(*Appendixes continue*)

The connections linking EC_in to CA1 and CA1 to EC_out are not modified in the course of the simulated memory experiment. Rather, we pretrain these connections so they form an invertible mapping, whereby the CA1 representation resulting from a given EC_in pattern is capable of recreating that same pattern on EC_out. CA1 is arranged into eight columns (consisting of 80 units apiece); each column receives input from three slots in EC_in and projects back to the corresponding three slots in EC_out. See O'Reilly and Rudy (2001) for discussion of why CA1 is structured in columns.

Lastly, our model incorporates the claim, set forth by Michael Hasselmo and his colleagues (see, e.g., Hasselmo & Wyble, 1997), that the hippocampus has two functional modes: an encoding mode, where CA1 activity is primarily driven by EC_in, and a retrieval mode, where CA1 activity is primarily driven by stored memory traces in CA3; recently, Hasselmo, Bodelon, and Wyble (2002) presented evidence that these two modes are linked to different phases of the hippocampal theta rhythm. Although we find the theta-rhythm hypothesis to be compelling, we decided to implement the two modes in a much simpler way—specifically, we set the scaling factor for the EC_in to CA1 projection to a large value (6) at study, and we set the scaling factor to zero at test. This manipulation captures the essential difference between the two modes without adding unnecessary complexity to the model.

## Appendix C

## Basic Parameters

Table C1
*Basic Parameters for the Hippocampal and Cortical Models*

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $E_l$ | 0.15 | $g_l$ | 0.235 |
| $E_i$ | 0.15 | $g_i$ | 1.0 |
| $E_e$ | 1.00 | $g_e$ | 1.0 |
| $V_{rest}$ | 0.15 | $\Theta$ | 0.25 |
| $\tau$ | .02 | $\gamma$ | 600 |
| MTLC $\varepsilon$ | .004 | Hippo $\varepsilon$ | .01 |
| MTLC savg_cor | .4 | Hippo savg_cor | 1 |

*Note.* MTLC = medial temporal lobe cortex; Hippo = hippocampal; savg_cor = correction for sending layer average activation.

Twenty items at study: 10 target items (which are tested) followed by 10 interference items (which are not tested).

Twenty-percent overlap between input patterns (flip 16/24 slots).

Fixed high recall criterion, recall = .40.

Table C1 shows the other basic parameters, most of which are standard default parameter values for the Leabra algorithm.

For those interested in exploring the model in more detail, it can be obtained from Kenneth A. Norman's Computational Memory Laboratory web site: http://compmem.princeton.edu.

Received August 13, 2001
Revision received July 25, 2002
Accepted August 7, 2002 ■