Shimamura, A. P., Jurica, P. J., Mangels, J. A., Gershberg, F. B., & Knight, R. T. (1995). Susceptibility to memory interference effects following frontal lobe damage: Findings from tests of paired-associate learning. *Journal of Cognitive Neuroscience, 7,* 144–152.

Shimamura, A. P., Jernigan, T. L., & Squire, L. R. (1988). Korsakoff's Syndrome: Radiological (CT) findings and neuropsychological correlates. *Journal of Neuroscience, 8,* 4400–4410.

Shimamura, A. P., & Squire, L. R. (1986a). Korsakoff syndrome: A study of the relation between anterograde amnesia and remote memory impairment. *Behavioral Neuroscience, 100,* 165–170.

Shimamura, A. P., & Squire, L. R. (1986b). Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 452–460.

Shimamura, A. P., & Squire, L. R. (1988). Long-term memory in amnesia: Cued recall, recognition memory, and confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 763–770.

Smith, M. L., & Milner, B. (1984). Differential effects of frontal-lobe lesions on cognitive estimation and spatial memory. *Neuropsychologia, 22,* 697–705.

Squire, L. R. (1987). *Memory and brain.* New York: Oxford University Press.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231.

Squire, L. R., Amaral, D. G., & Press, G. A. (1990). Magnetic resonance measurements of hippocampal formation and mammillary nuclei distinguishes medial temporal lobe and diencephalic amnesia. *Journal of Neuroscience, 10,* 3106–3117.

Squire, L. R., Haist, F., & Shimamura, A. P. (1989). The neurology of memory: Quantitative assessment of retrograde amnesia in two groups of amnesic patients. *Journal of Neuroscience, 9,* 828–839.

Squire, L. R., & Zouzounis, J. A. (1988). Self-ratings of memory dysfunction: Different findings in depression and amnesia. *Journal of Clinical and Experimental Neuropsychology, 10,* 727–738.

Chapter 12

McClelland, J. L. (1996). In L. Reder, Implicit memory and
metacognition: Mahwah, NJ: Erlbaum, 275-286

Neural Mechanisms for the Control and Monitoring of Memory: A Parallel Distributed Processing Perspective

James L. McClelland
Carnegie Mellon University and the
Center for the Neural Basis of Cognition

The chapters by Shimamura (11), Norman and Schacter (10) and Funnell, Metcalfe, and Tsapkini (7) in this volume exhibit a striking integration of neuroscientific and psychological approaches to metacognition. It seems evident that the boundaries that once existed between these approaches are in the process of vanishing completely. Behavioral, functional imaging, and neuropsychological investigations are providing converging constraints. These constraints are leading to the emergence of theories of (a) the mechanisms of control of information processing; (b) the mechanisms of explicit and implicit memory; and (c) the interplay of these mechanisms in the monitoring and control of memory.

Several modeling frameworks have been developed in which theories relevant to these matters can be cast. Metcalfe's CHARM model (Metcalfe, 1990, 1993; Metcalfe, Cottrell, & Mencl, 1993) and Anderson's ACT framework (Anderson, 1983, 1993; Kimberg & Farah, 1993) have both been used productively in efforts to understand aspects of control, memory, and metacognition–cognition. Other theories have been developed within the context of the Parallel Distributed Processing (PDP) framework (Rumelhart, McClelland, & the PDP Research Group, 1986; McClelland, Rumelhart, & the PDP Research Group, 1986), as this has continued to evolve in research at Carnegie Mellon. This framework provides the prospect of an eventual bridge between the neural and cognitive levels of description, because the constructs used in these models—the units and connections—have a relatively direct mapping to the elements found in the brain—the neurons and synapses. For practical reasons, actual models generally simulate very large populations of neurons with small or moderate numbers of simulated units, but there is an explicit effort to understand the relationship between these two levels of modeling and to incorporate known aspects of the physiology into the modeling effort. Thus, aspects of the physiology can influence model details in ways that are relatively direct and explicit compared to the other approaches.

In this commentary, I consider the relationship between the concepts introduced in these theories and the ideas and findings reported in the three chapters that are the subjects of this commentary. First I review existing models relevant to the aforementioned points a–c, with primary focus on the theories that have been developed within the PDP framework. Then I consider how they might address, or be extended to address, the phenomena introduced in the chapters by Shimamura, by Norman and Schacter, and by Funnell et al. I also relate the interpretations offered by the PDP models to those offered in these three chapters.

CONTROL OF PROCESSING

A recent series of papers develop a PDP model of the mechanisms through which cognitive processes can be controlled. The essential idea was first presented in Cohen, Dunbar, and McClelland (1990) and has since been elaborated and extended in several other publications (Cohen & Huston, 1994; Cohen & Servan-Schreiber, 1992; Cohen, Servan-Schreiber, & McClelland, 1992). The initial version of the model from Cohen et al. (1990) illustrates the basic theoretical proposal (see Fig. 12.1). The idea is that information processing takes place in processing pathways consisting of interconnected modules, subject to control by activity in another module which we might call the "task" module.

The connection weights among the units in the pathways come through practice to encode stimulus–response relationships, so that the presentation of an appropriate stimulus at the input end of one of these pathways tends to lead to the activation of a corresponding response at the output end. The model was applied to the specific case of the Stroop tasks, in which words are printed in colored ink, and the task is to either name the color of the ink or read the word. To model these tasks, two pathways are introduced—one for processing the color of the ink, and one for processing the identity of the word; these converge on the same response outputs, so that both pathways can contribute to the determination of a response. Connections in these pathways are set by a learning rule; more training on word reading than color naming translates into stronger connections in the word reading pathway than the color-naming pathway, making the former stronger than the latter. The processing is cascaded (McClelland, 1979), so that activations build up gradually over time; responses are emitted when the activation of one of the output units reaches a fixed threshold. Stronger pathways tend to lead to faster responses, and the speedup produced by practice tends to follow a power law.

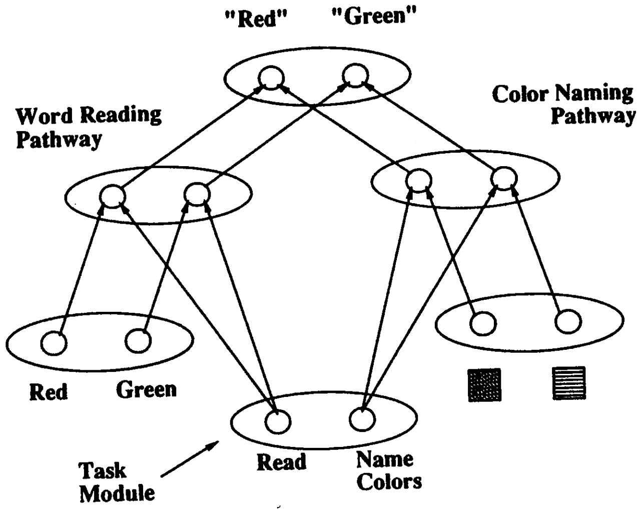

FIG. 12.1. (a) A sketch of the PDP model of control of processing in the Stroop task proposed by Cohen et al. (1990). Colored inputs corresponding to Red and Green are indicated with diagonal and horizontal stripes, respectively. © 1990 by the American Psychological Association. Adapted with permission.

In the absence of influence from the task module, each pathway may be capable of producing a correct output when an appropriate stimulus is presented to it, and when color and word are pitted against each other (e.g., the word RED is shown in green ink), the stronger word reading path will dominate. The task module contains units that serve to modulate processing in the two pathways, so that the intrinsically weaker pathway can dominate performance on conflict trials when the task is color naming. In the model as implemented, there are individual units for representing each task; for example, when the task is color naming, the color task unit is activated. This task unit has connections that allow it to prime the units in the intermediate layer of the color-naming pathway while at the same time inhibiting the intermediate layer of the word reading pathway; this facilitates processing in the color-naming pathway and suppresses processing in the word-reading pathway so that the color pathway can determine which response unit will reach threshold first. The strong-word pathway still exerts some influence, thereby accounting for the fact that there are interference and facilitation effects of word identity, when the task is color naming. When the task is word reading, the combined effect of task-based priming and the relatively greater intrinsic strength of the word-reading pathway are such that the effect of the color-naming pathway is overwhelmed, thereby accounting for the fact that words influence color naming but colors do not influence word reading.

The aforementioned approach places control of processing in the "task units"; these units do not actually carry out the task, but they influence units elsewhere that do. Deficits in the activation of these units would, therefore, be expected to lead to deficits in the control of processing. This idea is the basis for the model of schizophrenic deficits in the control of processing offered by Cohen and Servan-Schreiber (1992). They proposed that the task modules necessary for control of behavior are located in the prefrontal cortex, and that a schizophrenia produces a deficit in dopamine systems that project to the prefrontal cortex, giving rise to a reduction of responsivity of prefrontal neurons involved in maintaining task representations, and a corresponding reduction in control of behavior. They were able to use this model to account for several aspects of schizophrenic deficits in information processing tasks, including a reduction in the ability to inhibit prepotent responses to stimuli and override them with alternative responses consistent with task instructions, as in the Stroop task. Cohen and Servan-Schreiber also generalized the notion of the task representation considerably, to accommodate the fact that schizophrenic deficits appear to apply not just to the control of behavior by the general form of the task, but also to the influence on responding of contextual information. For example, in one experiment, schizophrenics were impaired in the use of context to select a relatively infrequent meaning of an ambiguous word (e.g., PEN, as fenced enclosure rather than writing implement). They argued that the frontal lobes play a crucial role in maintaining representations of context—whether generalized task context or specific local context—in a form that allows it to influence information processing, and their simulations encompassed deficits in use of both task context and local language context.

As originally formulated, the model of Cohen and Servan-Schreiber (1992) does not provide any account of more complex tasks that depend on frontal functions, such as motor sequencing or the Wisconsin Card sort task. These tasks have, however, been addressed in other models of frontal functioning. Indeed, Kimberg and Farah (1993) simulated both of these tasks using a production system model based on Anderson's ACT–R framework (Anderson, 1993). Production system models are useful in such cases, because the implementation of sequential processes in connectionist models can be cumbersome, and the graded strengths and activations provided by contemporary production system models allows them to capture what for present purposes may be the crucial characteristics of these sequential processes as they would arise in a connectionist system. In any case, Kimberg and Farah (1993) simulated performance of normals and frontal patients in the motor sequencing and Wisconsin Card Sort tasks (as well as a memory task to be considered here later). To simulate the effects of frontal lesions, they assumed that frontal damage weakens associations among elements in working memory. In many ways the model is quite similar to the model of Cohen and Servan-Schreiber, and it seems very likely that a version of this model could be developed even more in keeping with the assumptions of Cohen and Servan-Schreiber, if it was reformulated so that effects of damage weakened the activations of elements in working memory, rather than the associations between these elements. In the following I proceed as if this point were established, with the proviso that it does need checking through simulations.[1]

MEMORY

To apply the ideas embodied in the models discussed earlier to the domain of memory, we need a brain-systems model of the organization of memory function. Such a model has recently been formulated in the connectionist framework by McClelland, McNaughton, and O'Reilly (1995). A sketch of the overall structure of the model is shown in Fig. 12.2. This model shares with the models already discussed the idea that information processing takes place through the propagation of activation via processing pathways. Processing originates with peripheral sensory pathways that transduce inputs to higher level perceptual areas that are interconnected with other higher level areas, including higher level movement or response-planning areas that can propagate activation to muscles to produce overt responses, via peripheral motor pathways. The pathways of the Stroop model just described are like the pathways envisioned for this system, and the assumption that the skill of color-word reading is mediated by the strengths of connections in the word reading pathway still applies: This skill is thought of as reflecting the gradual accumulation of small changes to connections within the processing pathways that take place on every learning trial. In this model, implicit learning phenomena reflect these small changes (as well as other possible aftereffects, such as slight changes in the thresholds for activations of neurons in these pathways that have recently been activated). The small changes, however, that mediate implicit learning are not

[1] It seems likely that the specific assumptions Kimberg and Farah (1993) made about the nature of the information in working memory and the contents of the actual productions used to carry out task performance would have to be adjusted before the reformulated model could be successful.

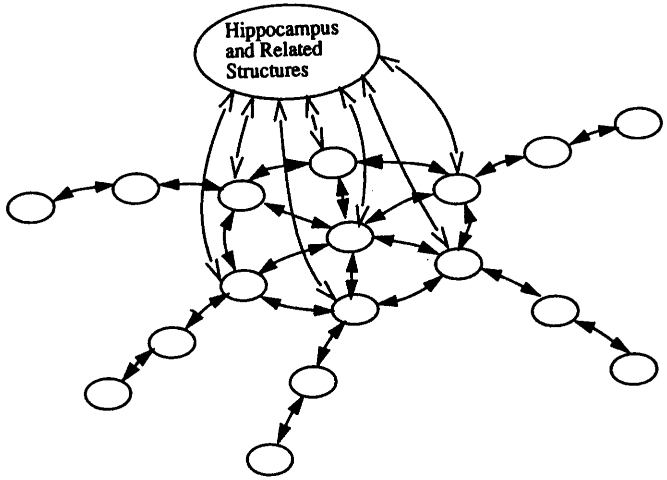

FIG. 12.2. A sketch of the brain systems model of information processing and memory proposed by McClelland et al. (1995). See text for further discussion. From McClelland (1994). Reprinted with permission of the publisher.

thought to be sufficient to support the rapid formation of new explicit memories. The reason for this is that explicit memories require the formation of associative links between arbitrary elements, and the formation of such associations requires much larger changes to the strengths of connection weights than those that are sufficient to produce implicit learning. McClelland et al. (1995) argued that it is crucial to the successful development of cortical processing systems that they learn slowly; if the large connection-weight changes necessary to form associative links between arbitrary materials were made quickly, they would lead to catastrophic interference with the structured knowledge and skills that are gradually built up in the cortex through the gradual accumulation of small changes (see McClelland et al., 1995, for further discussion).

In any case, McClelland et al. (1995) have argued that the hippocampus and related structures in the medial temporal lobes collectively called "the hippocampal system" provide a second learning system optimized to allow the rapid formation of associations between arbitrary elements. Such associations are assumed to underlie episodic and semantic memories in the form in which they are initially acquired. Loss of the hippocampal system therefore prevents the formation of new semantic and episodic memories and removes the substrate of recently formed semantic and episodic memories, thereby accounting for the anterograde and retrograde amnesia for such material found in patients with extensive lesions to these structures. Through repetition, however, both episodic and semantic memories can become integrated directly into the connections among units in the cortical processing pathways. This integration process accounts for the fact that retrograde amnesia for both semantic and episodic material is temporally limited in scope after extensive damage to the hippocampal system.

Before turning to the role of the frontal lobes in memory, it will be useful to characterize in a bit more detail the processes McClelland et al. (1995) assume take place during the storage and later retrieval of a new episodic memory, and to relate these to aspects of the memory models developed in other frameworks. Experiencing an episode, in the McClelland et al. model, produces a pattern of activity that is widely distributed over the higher level areas of the cortical processing system, including all (attended) aspects of the experience and any elaborations and associations to this experience that arise at the time. For example, an individual might hear the words "locomotive–dishtowel" paired together during an experimental session, and might form an elaborated representation involving a dishtowel wrapped around a (toy) locomotive, perhaps held in the hands of a small boy drying the locomotive after dropping it in a sink by accident. Pathways from these cortical areas into the hippocampal system produce a reduced description of the entire pattern. The elements of this reduced description are then bound together by associative synaptic modification within the hippocampal system. Later, an opportunity arises for recall of the learned episode whenever a part of the experience becomes reactivated in the neocortex. For example, an experimenter might remind the individual of an earlier session in which a series of word pairs were presented, and then ask the subject to recall the second member of each word pair, after presentation of the first. The subject might then use the reminder of the earlier session as a retrieval cue to recall the general context of the session, and might hold this in mind as the probe words are presented. When "locomotive" is presented, the context together with the probe would be used together to probe the memory system. A neocortical representation containing these two would be projected to the hippocampal system, where it would produce a reduced description that would serve as a probe of the associative memory. The associative synaptic modifications that took place during storage would then lead to the completion of the rest of the pattern. Return connections from the hippocampal system to the neocortex would then provide the basis for the reinstatement of a pale replica of the entire neocortical pattern that had been present during study, which (in successful cases) would provide enough of a reinstatement of the locomotive–dishtowel representation to allow for the readout of an appropriate overt response.

Perhaps it is worth considering briefly the relation of this model to other models, such as Metcalfe's CHARM and Anderson's ACT. All of the models view retrieval as dependent on associative connections and all provide mechanisms in which the memory probe plays a crucial role (see also Humphreys, Bain, & Pike, 1989). CHARM adds a notion of familiarity monitoring not explicitly used in McClelland et al. (1995), but that is consistent with many physiological findings. Most of the physiological work suggests that the actual computation of familiarity is carried out in the medial temporal lobes; this does not necessarily contradict Metcalfe's suggestion that a deficit in monitoring familiarity in frontal lobes could account for some frontal deficits, because the results of the computation might well be propagated to the PFC, for use in the control of processing.

CONTROL OF MEMORY: MAINTENANCE OF CONTEXT

Given all of the previous discussion, we are now in a position to consider what role the frontal lobes might play in various memory tasks. The natural suggestion is that they provide the mechanisms responsible for the control of memory; specifically, those neural structures responsible for the formation and maintenance of contextual representations that are normally part of the input to the hippocampal memory system during learning and during test. Such representations can be thought of as being among the contents of the active or working memory of the subject.

This general hypothesis, which I will call the *maintenance of context* hypothesis, needs further elaboration before it will serve as even a useful starting place for the development of a model of the role of the frontal lobes in memory retrieval. In the following section I review the relevant data and suggest some directions that might be followed toward the formulation of these elaborations.

MEMORY DEFICITS RESULTING FROM FRONTAL LESIONS

We can now consider the pattern of deficits seen in frontal patients in light of the hypothesis just stated. Simulations will ultimately be necessary to substantiate the claims I make in this section, but in most cases the claims can be supported by pointing to relevant characteristics of existing simulation models that illustrate the ideas in other contexts.

As an initial phenomenon to be explained, we can note that frontal patients generally show disproportionate deficits in recall relative to recognition; that the recall deficit is most pronounced in free recall; and that the deficit is substantially reduced if recall cues to individual memory traces are provided. Shimamura (Chapter 11, this volume) provides extensive discussion of these points. He suggests that the deficit may lie in the ability to formulate and maintain internal representations of retrieval cues sufficient to focus the retrieval process on the desired memory target.

These proposals are, of course, virtually equivalent to the maintenance of context hypothesis. Once the hypothesis is formulated within an explicit computational framework, its implications for processing can be directly assessed. Although such a model for memory retrieval remains to be developed, some of these implications can perhaps be anticipated, based on observations of the effects of degradation of context representations in the model of Cohen and Servan-Schreiber (1992). Their simulations illustrated that degradation of context representations tended to produce two complementary effects: first, degradation reduced the ability to produce *less* typical or frequent responses; second, it led to an increase in the tendency to intrude *more* typical or frequent responses. Cohen and Servan-Schreiber used this finding to argue that both types of findings could reflect a single underlying deficit. In this light it is interesting to note that frontal patients often intrude strong preexisting associates, overriding newly acquired, context-specific associations in memory experiments. By extension of the proposal of Cohen and Servan-Schreiber, we might suggest that both poorer overall recall, and the intrusion of preexisting associates results from the failure to form an adequate representation of the context that restricts recall to the desired target.

However, there are two very different ways of thinking about the reasons why frontal patients might have difficulty forming and maintaining appropriate representations of context. One way would be to assume that there is damage directly to the module or modules containing the units that actually represent the characteristics of the context. Such damage would be expected to degrade the context representation, leading it to serve as a relatively poor retrieval cue. Alternatively, one could assume that other modules that support the formation and maintenance of the context representation are affected.

It seems likely that something more like the latter is involved, at least for some patients. One reason for this is the observation that if information about the context is supplied as part of the memory probe, some frontal patients can show very mild deficits. Problems arise when the patient must generate and maintain the context, at least in some frontal patients (perhaps those with lesions in the dorsolateral prefrontal cortex).

It is possible, however, that other portions of the frontal lobes are directly responsible for the representation of certain types of content; specifically, content related to personal history and episodic structure. The analysis of patient DRB by Damasio, Eslinger, Damasio, Van Hoesen, and Cornell (1985) raises this possibility. This patient had extensive damage to the basal forebrain, cingulate gyrus, and inferolateral temporal cortex, as well as the medial temporal lobes, and exhibited very slight residual memory of his own personal history; he confabulated wildly about his own recent episodic experiences. The picture is quite different from that presented by pure medial temporal lobe patients, in which remote memory about personal history and early episodic information is retained. Whereas Damasio et al. suggested that inferolateral temporal cortex may play a crucial role in the loss of episodic memory, recent cases reported by Hodges, Graham, and Patterson (1995) suggested that lesions restricted to these areas produce a profound deficit of the semantic content of episodes but leave the memory for the event itself intact (the patient confuses similar objects within events but clearly remembers particular events subject to such confusions). Thus, the possibility remains that a portion of the frontal lobes, such as the basal forebrain, may play a crucial role in the representation of self-in-context. If so, disorders in this area would lead to gross disturbances in the ability to probe memory episodically, thus accounting for disturbances of some aspects of context memory in some frontal patients.

These last comments dovetail with Norman and Schacter's ideas about differential effects of different types of frontal lesions. They suggest that some types of frontal lesions may lead to a degradation of representations of context, either by eliminating some aspects of the context representation, or by making the activation of such information very diffuse. This latter possibility is certainly consistent with the notion that some portion of the frontal lobes might actually provide the representation of self in relation to context. The possibility that severe degradation of these representations might be the basis for memory confabulation deserves careful exploration.

Obviously the ideas just presented are highly speculative. There is an obvious need to elaborate a neuropsychological model of the retrieval process and to specify how the aspects of the model relate to various regions of the brain in order to address these and other aspects of the data.

These ideas are also quite incomplete. The monitoring of memory is clearly an important function, and there seems to be general consensus that frontal damage disturbs these monitoring functions, but I have hardly touched on this matter in these comments. One possible way of relating the maintenance of context hypothesis to failures of memory monitoring might be to suggest that, at least in some cases, it may not be the monitoring per se that is at fault, but the ability to maintain representations that reflect the results of monitoring, so that they can be used to guide future behavior. This idea is consistent with the observation of preservation errors in tasks like the Wisconsin Card Sort, where the patient is given clear feedback and is clearly aware of the feedback at some level, but is unable to use it to guide his behavior.

Obviously, then, we remain a long way from a full characterization of the neural basis of the control and monitoring of memory. We can expect rapid growth in our empirical understanding of the role played by various parts of the frontal lobes, and other brain areas, in memory and the control of memory through the continued use of neuropsychological and functional imaging approaches. Additional computational modeling work should help clarify and integrate the results of these empirical investigations.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the stroop effect. *Psychological Review, 97,* 332–361.

Cohen, J. D., & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. In C. Umilta & M. Moscovitch (Eds.), *Attention & performance xv: Conscious and nonconscious information processing* (pp. 453–476). Cambridge, MA: MIT Press.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99,* 45–77.

Cohen, J. D., Servan-Schreiber, D., & McClelland, J. L. (1992). A parallel distributed processing approach to automaticity. *American Journal of Psychology, 105,* 239–269.

Damasio, A. R., Eslinger, P. J., Damasio, H., Van Hoesen, G. W., & Cornell, S. (1985). Multimodal amnesic syndrome following bilateral temporal and basal forebrain damage. *Archives of Neurology, 42,* 252–259.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory, 3,* 463–495.

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.

Kimberg, D. Y., & Farah, M. J. (1993). A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior. *Journal of Experimental Psychology: General, 122,* 411–428.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

McClelland, J. L. (1994). The organization of memory: A parallel distributed processing perspective. *Rev. Neurol.* (Paris), *150,* 570–579.

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.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119,* 145–160.

Metcalfe, J. (1993). Novelty monitoring, metacognition and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3–22.