Laminar Cortical Dynamics of Cognitive and Motor Working Memory, Sequence Learning and Performance: Toward a Unified Theory of How the Cerebral Cortex Works

Stephen Grossberg[1] and Lance R. Pearson[2]*

Department of Cognitive and Neural Systems
Center for Adaptive Systems[1]
and
Center of Excellence for Learning in Education, Science, and Technology
677 Beacon Street
Boston University
Boston, MA 02215
617-353-7858 (Phone)
617-353-7755 (Fax)
steve@bu.edu
http://www.cns.bu.edu/Profiles/Grossberg

Department of Psychology[2]
Vanderbilt University
111 21st Avenue South
Nashville, TN 37203
502-741-6493 (Phone)
LancePearson@gmail.com

Corresponding Author: Stephen Grossberg

Technical Report CAS/CNS TR-08-002

Submitted: April, 2006

Running Title: Cortical Working Memory and Sequence Learning

Key words: working memory, competitive queuing, immediate serial recall, immediate free recall, delayed free recall, continuous-distracter free recall, sensory-motor imitation, chunking, sequence learning, prefrontal cortex, parietal cortex, position coding, rank order cells, cerebral cortex, laminar computing

* Authors are listed in alphabetical order.

[1] Supported in part by the National Science Foundation (NSF SBE-0354378) and the Office of Naval Research (ONR N00014-01-1-0624).

[2] Supported in part by the Air Force Office of Scientific Research (AFOSR F49620-01-1-0397), the National Science Foundation (NSF SBE-0354378), and the Office of Naval Research (ONR N00014-95-1-0409 and ONR N00014-01-1-0624). Present address: Department of Psychology, Vanderbilt University, 111 21st Avenue South, Nashville, TN 37203

ABSTRACT

How does the brain carry out working memory storage, categorization, and voluntary performance of event sequences? The LIST PARSE neural model proposes an answer to this question that unifies the explanation of cognitive, neurophysiological, and anatomical data from humans and monkeys. It quantitatively simulates human cognitive data about immediate serial recall and free recall, and monkey neurophysiological data from the prefrontal cortex obtained during sequential sensory-motor imitation and planned performance. The model clarifies why both spatial and non-spatial working memories share the same type of circuit design. It proposes how the laminar circuits of lateral prefrontal cortex carry out working memory storage of event sequences within layers 6 and 4, how these event sequences are unitized through learning into list chunks within layer 2/3, and how these stored sequences can be recalled at variable rates that are under volitional control by the basal ganglia. These laminar prefrontal circuits are variations of laminar circuits in the visual cortex that have been used to explain data about how the brain sees. These examples from visual and prefrontal cortex illustrate how laminar neocortex can represent both spatial and temporal information, and open the way towards understanding how other behaviors may be represented and controlled by variations on a shared laminar neocortical design.
.

1. Introduction. Intelligent behavior depends upon the capacity to think about, plan, execute, and evaluate sequences of events. Whether we learn to understand and speak a language, solve a mathematics problem, cook an elaborate meal, or merely dial a phone number, multiple events in a specific temporal order must somehow be stored temporarily in working memory. As event sequences are temporarily stored, they are grouped, or chunked, through learning into unitized plans, and can later be performed at variable rates under volitional control either via imitation or from a previously learned plan. How these processes work remains one of the most important problems confronting cognitive scientists and neuroscientists (e.g., Cooper & Shallice, 2000).

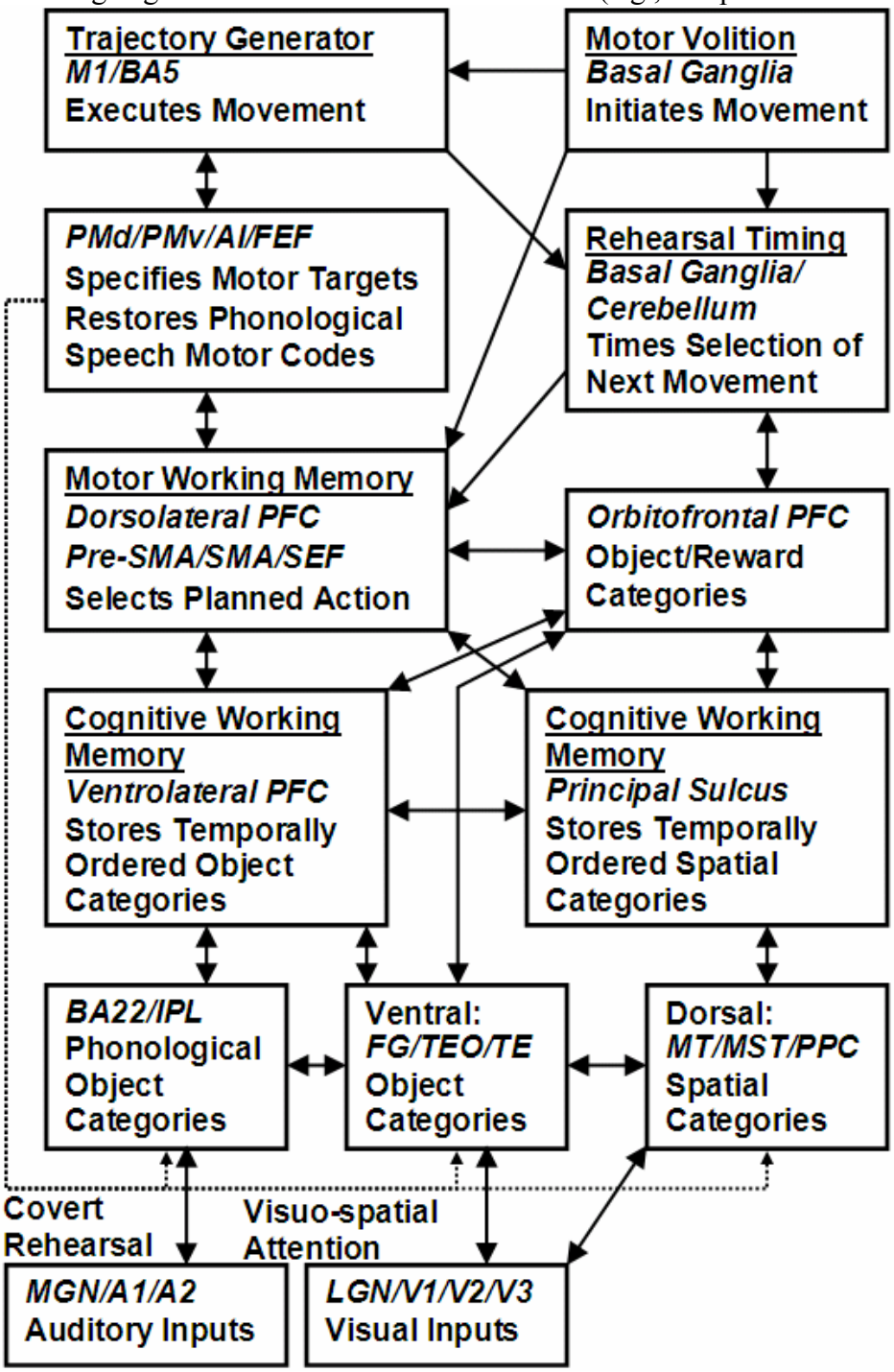


Figure 1. Schematic of the LIST PARSE model with proposed regional localizations in italics. The proposed function of each region is shown in plain text and names of components that are modeled in this paper are underlined. Model connectivity has been simplified. A more detailed specification is provided in Sections 4 and 5 and in Figure 2. Abbreviations: MGN, thalamic medial geniculate nucleus; A1/A2, core and extended auditory cortex in the superior temporal gyrus; LGN, thalamic lateral geniculate nucleus; V1/V2/V3, early visual processing occurring in the occipital lobe; BA22, Brodmann's area 22 in the superior temporal gyrus, corresponding to Wernicke's Area in the left hemisphere; IPL, areas in the inferior lateral parietal lobule, including the posterior sylvian fissure near the temporal-parietal boundary (area Spt); Ventral: FG/TEO/TE, portions of the ventral "what" processing stream in the inferotemporal (IT) cortex including areas of the Fusiform Gyrus; Dorsal: MT/MST/PPC, portions of the dorsal "where" processing stream including the posterior middle temporal gyrus and the posterior parietal cortex, including Brodmann's areas 7a,c,ip,m; VLPFC, ventrolateral prefrontal cortex (specifically areas 44, 45 and 47/12); Principal Sulcus, portions of the dorsolateral prefrontal cortex corresponding to the depths of the principal sulcus in the macaque monkey (portions of areas 46, 9/46v,d); Dorsolateral PFC, the remainder of the dorsolateral prefrontal cortex (8, 9, and portions of 46, 9/46v/d); pre-SMA/SMA, pre-Supplementary Motor Area and Supplementary Motor Area; SEF, Supplementary Eye Field; Orbitofrontal PFC, orbitofrontal prefrontal cortex; PMd/PMv, dorsal and ventral portions of the lateral premotor cortex; AI, anterior insula; FEF, Frontal Eye Field; M1, primary motor cortex; BA5, Brodmann's area 5 in the parietal lobe.

This article introduces the LIST PARSE (Laminar Integrated Storage of Temporal Patterns for Associative Retrieval, Sequencing and Execution) model. LIST PARSE (Figures 1 and 2) proposes how the layered circuits of prefrontal and motor cortex are organized to realize sequential working memory storage, categorization, and motor planning and execution during both cognitive and sensory-motor tasks. The model makes predictions about the laminar organization of prefrontal cortical circuits that go beyond present neurophysiological and anatomical knowledge. It formulates these predictions by integrating and explaining several types of data and theoretical constraints: behavioral and neuroimaging data about cognitive information processing in humans; behavioral and neurobiological data about sensory-motor storage and performance of sequential actions in monkeys; anatomical data about the laminar prefrontal cortex; non-laminar models of cognitive and sensory-motor processing that have been used to explain and simulate behavioral and neurobiological data in these domains; and laminar models of visual cortex that have quantitatively simulated perceptual and neurobiological data about visual perception. These multiple experimental and theoretical constraints rule out many model hypotheses and mechanisms that might otherwise seem plausible when only confronted by one type of constraint.

LIST PARSE predicts how variations on laminar cortical circuits that have explained data about visual perception may also explain data about cognitive information processing. The article hereby advances efforts to develop a unified theory of neocortex wherein many different types of behavior may all be derived as emergent properties of a shared laminar cortical design. This new paradigm of Laminar Computing promises to revolutionize our understanding of intelligent computation (Grossberg, 2003a). In this paradigm, the deep layers (layers 6-4) of granular neocortex are proposed to carry out *filtering* and *temporary storage* of incoming items

while maintaining sensitivity to item activation differences (*normalization*) and selectively emphasizing relevant features of the input pattern (*contrast enhancement*). Superficial cortical layers (layers 2/3) are proposed to *group* items represented in the deeper layers (Figure 2). For example, whereas the superficial layers in visual cortex group inputs based on properties like stimulus colinearity (Grossberg, 1984; Grossberg & Mingolla, 1985; Grossberg, Mingolla, & Ross, 1997; Kapadia, Ito, Gilbert, & Westheimer, 1995; von der Heydt, Peterhans, & Baumgartner, 1984), the superficial layers of lateral prefrontal cortex are predicted to group based upon properties like sequential order. As this theory develops, results about the behaviors controlled by one neocortical area will provide increasingly strong constraints to guide the understanding of other, seemingly different, behaviors that are controlled by other neocortical areas. This theme should be of interest to all cognitive scientists and neuroscientists, as well as to designers of intelligent systems and the general public.

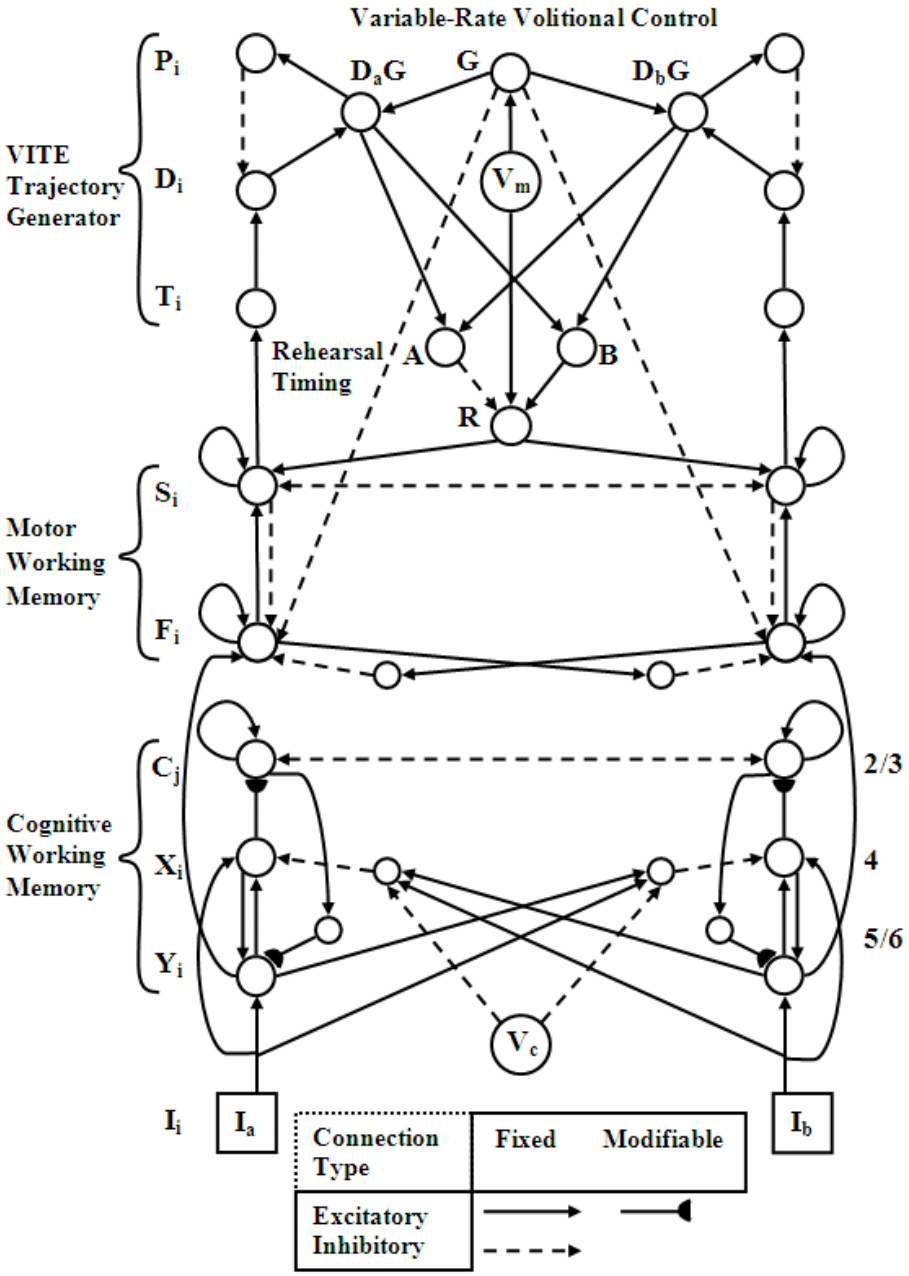


Figure 2. Circuit diagram of the LIST PARSE model. Solid arrows indicate fixed excitatory connections. Solid lines with hemi-disks indicate modifiable (i.e., learned) connections. Dashed arrows indicate fixed inhibitory connections. This convention is maintained throughout the remainder of the circuit diagrams. Only 1-item chunks (*C*) and their feedback connections (*M*) within a single cognitive working memory channel are shown, whereas the model uses chunks of various sizes in layer 2/3 and feedback from layer 2/3 to layer 5/6 of the cognitive working memory is broadly distributed. Also, only the excitatory projections from cognitive working memory to the motor plan field $(Y \rightarrow F)$ are shown. A more detailed treatment of these circuits is provided in Section 4.

LIST PARSE develops the proposal that all working memories are designed from similar circuits, whether to control free recall or serial recall in humans, or sequential read-out of planned movements in monkeys. The intellectual basis for this claim is the derivation by Grossberg (1978a) of all working memories whose temporarily stored sequence information can be chunked through learning and stably remembered in long-term memory. LIST PARSE shows how the use of shared working memory properties can explain different sorts of data properties. The model predicts how different experimental paradigms lead to different cognitive selection strategies, which result in different rehearsal and recall patterns. This theme should be of interest to all scientists who are interested in the temporal organization of behavior.

In particular, LIST PARSE proposes a unified account of cognitive data about both immediate serial recall and free recall. It quantitatively simulates bowing of serial position performance curves during serial recall (Cowan, Nugent, Elliott, Ponomarev & Saults, 1999; Henson, Norris, Page & Baddeley, 1996) and immediate and delayed free recall (Glanzer & Cunitz, 1966; Postman & Phillips, 1965; Tan & Ward, 2000), continuous-distracter free recall and aspects of long-term recency effects (Bjork & Whitten, 1974; Davelaar et al., 2005; Tan & Ward, 2000; Tzeng, 1973), error type distributions (Henson et al., 1996), list length effects (Baddeley, Thompson & Hitch, 1975; Crannell & Parrish, 1957), and temporal limitations upon recall (Murdock, 1961). The model also qualitatively explains effects of attention upon serial performance curves (Cowan, Nugent, Elliott, Ponomarev & Saults, 1999), temporal grouping effects (Hitch, Burgess, Towse & Culpin, 1996; Ryan, 1969), presentation variability (Knoedler, Hellwig & Neath, 1999; Tan & Ward, 2000), phonemic similarity effects (Baddeley, 1986), non-word lexicality (Besner & Davelaar, 1982), word frequency, item familiarity, and list strength effects (DeLosh & McDaniel, 1996; Ratcliff, Clark & Shiffrin, 1990), distracter protocols (Watkins, Neath & Sechler, 1989), and modality effects (Baddeley, 1986; Crowder, 1978). Tables 1 and 2 summarize some of these phenomena.

Previous cognitive models have tended to explain data about either free recall or serial recall, but not both. The unified explanation of free recall and serial recall data by LIST PARSE derives from its use of a shared working memory circuit for short-term storage and read-out of sequential information. Indeed, free recall and serial recall experimental paradigms are artificial environments that have only very recently been experienced by humans. There is no reason to believe that such paradigms drove the formation of qualitatively distinct storage, learning, and performance mechanisms on the time scale of brain evolution. Correspondingly, the LIST PARSE model traces performance differences between free recall and serial recall to variations in selection strategies that operate on stored information in working memory, which lead to different rehearsal and recall properties.

Table 1. Behavioral phenomena observed in both immediate serial recall and immediate free recall protocols for which the LIST PARSE model gives a quantitative account:

Bowing of serial position performance curves: For a presented list, the first few items have an advantage in recall (primacy effect), though they are likely not the most recently rehearsed in either protocol. The last item(s) tend to have advantages in recall (recency effect) although they are commonly recalled last especially in the ISR task and often not recalled first in the IFR task (e.g., object serial recall: Hulme, Roodenrys, Schweickert, Brown, Martin & Stuart, 1997; spatial serial recall: Jones, Farrand, Stuart and Morris, 1995; free recall: Tan & Ward, 2000).

Error type distributions: Item recall errors like omissions, insertions and repetitions, occur in both protocols. However, order errors largely consisting of transpositions of neighboring items tend to make up the majority of errors at span length in the ISR task while they are not typically regarded as errors in the IFR task (e.g., Henson, Norris, Page & Baddeley, 1996).

Word length and list length effects: Performance for lists with more items, or words that take longer to articulate (e.g., more syllables), tend to be more poorly recalled that lists with shorter words or fewer items (see Section 3 for discussion of current debate about word length effects). Additionally, the span (length of lists subjects can recall correctly) tends to scale linearly with articulation rate (e.g., Baddeley, 1986; Baddeley, Thompson, & Hitch, 1975; Cowan et al., 1992; Crannell & Parrish, 1957; Hulme et al., 1997).

Temporal limitations upon recall: Increasing retention interval without rehearsal decreases list performance. Representations of portions of unrehearsed lists of items at span length are rendered irrecoverable by delays of as little as two to three seconds under most protocols and stimulus types (e.g., Baddeley, 1986). However, memory for sub-span lists may persist to some extent for non-rehearsed periods far exceeding those intervals (Murdock, 1961).

Table 2. Behavioral phenomena observed in both immediate serial recall and immediate free recall protocols for which the LIST PARSE model gives a qualitative account:

Effects of attention: In tasks where attentional resources are diverted, there are enhanced recency effects (e.g., Cowan, Nugent, Elliott, Ponomarev & Saults, 1999).

Temporal grouping effects: Inserting pauses between groups of items creates primacy and recency gradients within groups (e.g., serial recall: Henson et al., 1996; Ryan, 1969).

Presentation variability: Faster presentation rates (reduced IOIs) result in enhanced recency effects and diminished primacy effects. Increasing the delay after list presentation promotes primacy at the expense of recency effects, even in cases where rehearsal is not likely to be a significant factor (e.g., Watkins, Neath & Sechler, 1989; Tan & Ward, 2000).

Phonemic similarity effect: Performance in recalling a list composed of items that are phonologically confusable, such as the rhyming letter names B, D, C, and G, is worse than for low-confusability stimuli (e.g., Baddeley, 1986).

Non-word lexicality: Maximum number of items recalled (span) increases for non-words that sound like words relative to those that do not, regardless of the use of articulatory suppression, techniques aimed at restricting the ability to phonologically rehearse a list (e.g., Besner & Davelaar, 1982).

Word frequency/item familiarity and list strength effects: Maximum number of items recalled (span) increases with item familiarity. With items of differing retention strength (familiarity), weak items are recalled better when they are in lists of strictly weak items, whereas strong items are recalled less well in pure-strong lists than in mixed lists (e.g., DeLosh & McDaniel, 1996; Ratcliff, Clark & Shiffrin, 1990).

Distracter protocols: Adding a distracter task at the end of list presentation greatly reduces the recency effect. However, also adding a distracter task between each word presentation (called the continuous-distracter protocol) largely reinstates the recency effect (Watkins, Neath & Sechler, 1989).

Modality effects: Auditory list presentation tends to produce slightly more extended recency portions of the serial performance curve than visual presentation when subvocal rehearsal is prevented (Baddeley, 1986). However, an irrelevant auditory item presented at the end of the list largely eliminates the modality difference (suffix effect: Crowder, 1978).

LIST PARSE also quantitatively simulates sensory-motor data about sequential processing. Brain mechanisms of sequential storage and recall have been clarified through electrophysiological recordings from lateral prefrontal cortex of monkeys performing a sequential movement task (Averbeck et al., 2002, 2003a, 2003b) and functional imaging studies of humans conducting various working memory tasks (e.g., Braver, Barch, Kelley, Buckner, Cohen, Miezin, Snyder, Ollinger, Akbudak, Conturo & Petersen, 2001; Chein & Fiez, 2001; D'Esposito, Aguirre, Zarahn, Ballard, Shin & Lease, 1998; Fletcher & Henson, 2001; Marshuetz, Smith, Jonides, DeGutis & Chenevert, 2000; Passingham, 1993; Paulesu, Frith & Frackowiak, 1993; Poeppel, 1996; Smith, Jonides, Marshuetz & Koeppe, 1998; Ranganath, Johnson & D'Esposito, 2003). LIST PARSE simulates monkey sensory-motor data with the same model that it uses to simulate the human cognitive data. It also leads to a new proposal for how to explain cognitive and neurophysiological data showing conjunctive coding of item, order, and list position in a list.

LIST PARSE advances a theoretical synthesis of several types of data that reflect shared neural mechanisms for sequential storage and learning. However, the model in its current form does not presume to explain all the experimental effects that have been discovered through decades of research in the several fields for which it proposes a unification. Model strengths and proposals for future development are given below. Section 2 reviews modeling concepts on which LIST PARSE builds. Section 3 summarizes cognitive and brain data for which LIST PARSE proposes explanations. Section 4 qualitatively explains model processes. Section 5 summarizes model explanations and simulations of behavioral and brain data. Section 6 proposes a model extension that provides the first biologically plausible account of how conjunctive coding of item, order, and list position can occur in the brain. Section 7 reviews and compares related models. Section 8 offers concluding remarks. The Appendix summarizes model equations and parameters.

2. Modeling Background. LIST PARSE builds upon earlier models of (1) how individual events are categorized, (2) how these events are attended due to their motivational salience, and (3) how sequences of these events are stored temporarily in working memory, unitized through learning into list chunks, and sequentially performed at variable rates under volitional control. Only the processes in (3) are further modeled in this article. All three sorts of contributions are reviewed, however, since they are needed to understand how the processes in (3) work:

2.1. Categorizing Individual Items: Persistence, Priming, and Attention. Individual items (events, objects) are categorized, or unitized, at a sensory cortical stage, such as temporal cortex. An "item" in LIST PARSE is a category that represents a spatial pattern of features occurring within a short time interval at a still earlier processing stage, such as visual or auditory cortex. See the Object Categories and Spatial Categories in Figure 1, which correspond to item inputs I_i in Figure 2. We review key mechanisms of item learning because the same categorization

processes are predicted to occur at higher cortical areas to control sequence learning. Indeed, as explained below, sequences of events occurring through time are recoded by LIST PARSE into temporally evolving spatial patterns of activity across prefrontal cortical cell populations. In other words, there is a temporal-to-spatial recoding within short-term working memory. See the Cognitive Working Memory in Figures 1 and 2, which is realized by processes X_i and Y_i in Figure 2 (see Appendix equations (2) and (3)). LIST PARSE models how each such spatial pattern in working memory may be unitized through learning into a sequence, or list, category. See the list categories C_j in Figure 2 (Appendix equation (5)).

Adaptive Resonance Theory, or ART, models clarify how such item and list categories may be incrementally learned in real time through interactions between several cortical regions (sensory, temporal, parietal, prefrontal), specific and nonspecific thalamic nuclei, and the hippocampal system. In particular, ART predicts how bottom-up and top-down learning are coordinated during item category learning (e.g., Bradski & Grossberg, 1995; Carpenter & Grossberg, 1991, 1993; Grossberg, 1978a, 1980, 1999a; Grossberg & Merrill, 1996; Grossberg & Versace, 2006; also see reviews in Engel, Fries & Singer, 2001 and Pollen, 1999). Bottom-up learning tunes the pathways that selectively activate the item categories. Top-down learning encodes expectations that can match, synchronize, and amplify the bottom-up distributed features that their categories learn to bind. Top-down expectations, acting alone, can also modulate, or prime, their target cells during a delay period before bottom-up inputs are received. This combination of bottom-up and top-down learning helps to stabilize the bottom-up category learning.

In particular, match/mismatch operations between top-down expectations and bottom-up feature patterns permit discrimination between familiar and novel events, can drive reset of an active category and its top-down expectation during a mismatch, and trigger a memory search, or hypothesis testing, that leads to selection and learning of a better-matching category. See Grossberg (2003a) and Grossberg & Seidman (2006) for recent reviews. See Grossberg & Versace (2006) for an analysis of how laminar cortical circuits in multiple cortical and thalamic regions can work together synchronously to learn individual *item* categories. Such item categories are the inputs to the working memory circuits such as those modeled by LIST PARSE.

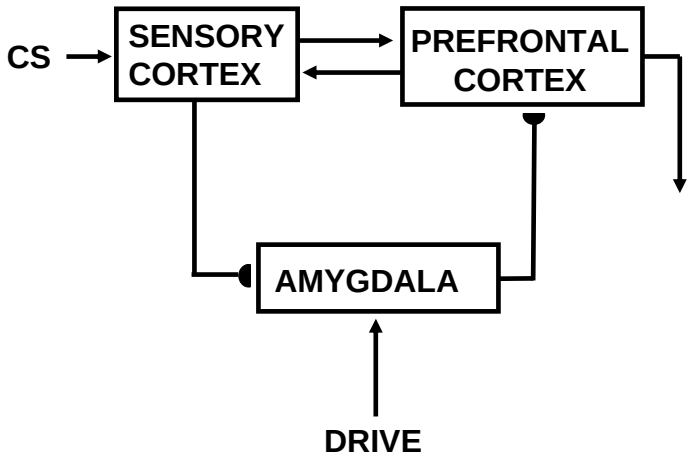
The bottom-up learning in LIST PARSE is learning of *list* categories (pathways $X_i \rightarrow C_j$ in Figure 2), whereas the top-down learning allows an already learned list category (C_j in Figure 2) to read-out into short-term working memory (pathways $C_j \rightarrow Y_i$ in Figure 2) the representation of the item sequence that the category represents. For a review of these process localizations in brain anatomical regions, see Supplementary Material, Section 1.

2.2. Motivationally Salient Item Categories and Adaptively Timed Delay Activity. Another line of modeling work analyzes cognitive-emotional interactions that help to select the item categories that will be stored in working memory. This work clarifies how item categories attract attention when they are motivationally salient, how motivated attention facilitates persistent storage of item categories in prefrontal cortex, and how attention and persistent storage are maintained during an adaptively timed delay interval.

The Cognitive-Emotional-Motor, or CogEM, model (e.g., Grossberg, 1971, 1982, 2000; Grossberg & Levine, 1987; Grossberg & Seidman, 2006) clarifies how item categories that are learned in the temporal cortex may become salient enough to influence working memory storage by the prefrontal cortex with motivational support from the amygdala. In particular, CogEM predicts how item categories in temporal cortex interact with orbitofrontal cortex and the amygdala, and how the amygdala interacts with orbitofrontal cortex (Figure 3). Reinforcement

learning enables item categories in the temporal cortex to activate the amygdala via learned *conditioned reinforcer* signals, and the amygdala to activate orbitofrontal projections of the item categories via learned *incentive motivational* signals. The conjoint action of temporal cortex and amygdala inputs to orbitofrontal cortex can activate the object/reward-sensitive cells that are found in orbitofrontal cortex (Rolls, 2000, 2004; Schoenbaum & Roesch, 2005; Schoenbaum & Setlow, 2001). In particular, learned incentive motivational signals from amygdala enhance the activation of task-relevant item category cells in orbitofrontal cortex and thereby facilitate storage of these item categories in working memory. Such enhanced orbitofrontal categories can, in turn, use top-down attentive expectations to selective amplify compatible data in temporal cortex (Barbas, Medalla, Alade, Suski, Zikopoulos & Lera, 2005).

(a)



(b)

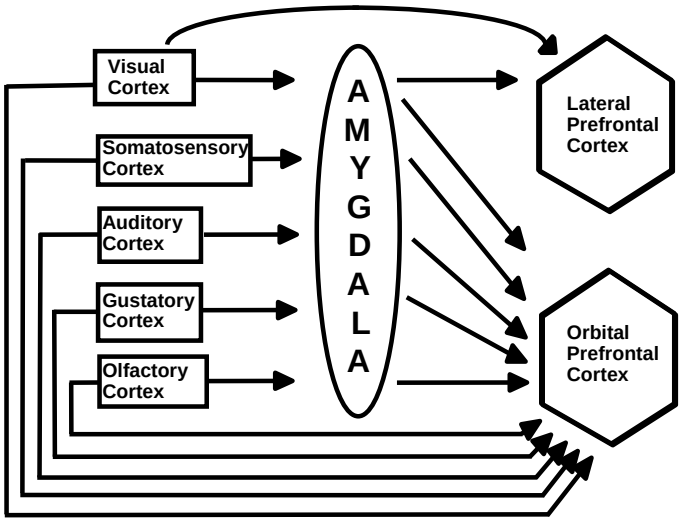


Figure 3. (a) The CogEM model describes how conditioned stimuli that activate a sensory cortex learn to activate reward-sensitive regions of the brain, such as the amygdala, through conditioned reinforcer learning. These conditioned stimuli also activate higher cortical areas, notably the orbitofrontal cortex. The prefrontal cells require motivational support from the amydgala to fire vigorously. This motivational support is enhanced by incentive motivational learning from the amygdala to the orbitofrontal cortex. Active orbitofrontal cells feed back attentive signals to the sensory cortices to select sensory representations that are consistent with the currently active drive state in the amygdala. As noted in Section 6, these orbitofrontal cells realize conjunctive coding of sensory and motivational information which may be compared with the conjunctive coding that Section 6 summarizes to explain positional coding in working memory. (b) An anatomical interpretation of the CogEM circuit. [Redrawn with permission from Barbas (2005).]

The START (Spectrally Timed ART) model (Grossberg & Merrill, 1992, 1996; Grossberg & Schmajuk, 1989; Grossberg & Seidman, 2006) extended the CogEM model to propose how motivated attention can maintain the activity of orbitofrontal item categories during a learned task-appropriate delay interval. This adaptively timed learning was proposed to occur within the dentate-CA3 region of the hippocampal system before projecting to orbitofrontal cortex. See Grossberg & Seidman (2006) for a review of both CogEM and START.

Many laboratories have reported that prefrontal cells can retain their activation during a delay interval after their eliciting sensory stimuli terminate (e.g., Curtis & D'Esposito, 2004; Curtis, Rao & D'Esposito, 2003; D'Esposito & Postle, 1999; Funahashi, Bruce & Goldman-Rakic, 1989, 1990, 1991; Funahashi, Inoue & Kubota, 1993, 1997; Fuster, 1973; Fuster & Alexander, 1971; Fuster, Bauer & Jervey, 1985; Kubota & Niki, 1971; Mechelli, Price, Friston & Ishal, 2004; Miller, Erickson & Desimone, 1996; Niki, 1974; Ranganath & D'Esposito, 2005; Stern, Sherman, Kirchhoff & Hasselmo, 2001). This delay period activity has sometimes been called "working memory".

This article adopts a more restrictive concept of working memory, and one that is consistent with psychological concept of a temporary workspace that supports mental manipulations. Here, the term working memory is only used to describe brain processes and regions that can simultaneously store the *temporal order of several events*. This distinction differentiates working memory from simpler short-term memory or persistence processes. Adaptively timed persistence of individual events, or sequential read-out from long-term memory (LTM), are not forms of working memory in our terminology. With this clarification in mind, the LIST PARSE model assumes that the inputs to working memory are unitized item categories that have enough adaptively timed, incentive motivational support to be persistently stored and transferred into working memory.

2.3. Working Memory Stores Temporal Order of Item Category Lists using Spatial Activity Gradients. As successive item categories are activated through time, they may sequentially store their item representations in working memory as a temporally evolving spatial pattern of activity across working memory cells (Figure 4). The *relative activity* of different cell populations codes the temporal order in which the items will be rehearsed. Items with the largest activities are rehearsed earliest (Grossberg, 1978a, 1978b). Hence, the original name *Item and Order working memory* for this class of models.

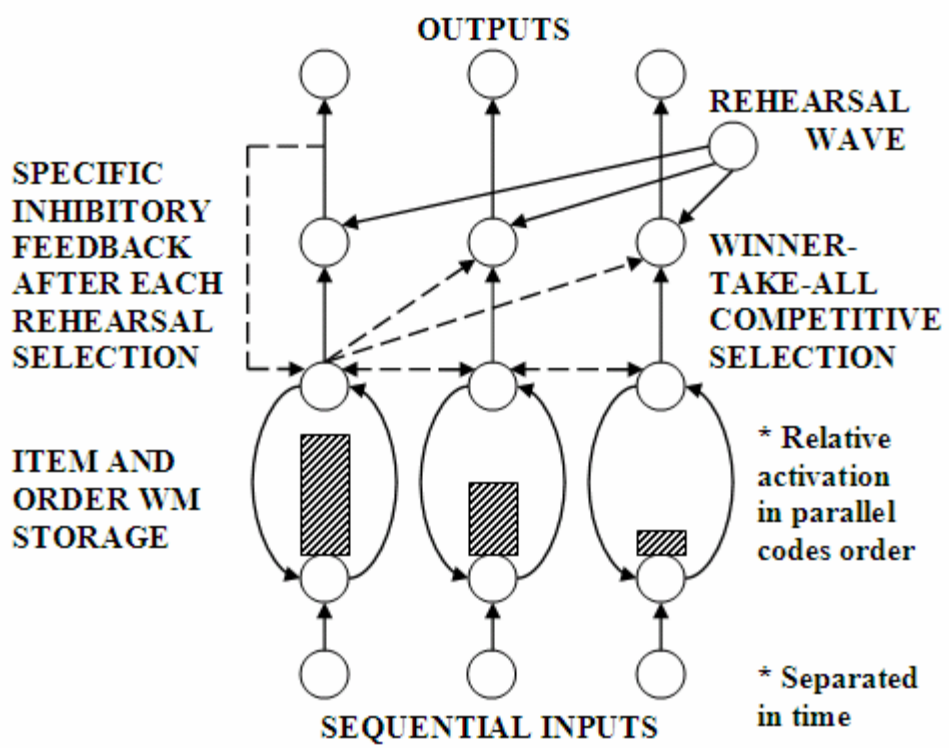


Figure 4. General framework for ordinal theories of serial order and competitive-queuing models of motor planning: temporal sequence of inputs creates a spatial activation pattern among STM activations, often a primacy gradient (height of hatched rectangles is proportional to cell activity). Relative activation level among items in STM codes for item and order. A rehearsal wave (global signal) allows item activations to compete before the maximal item elicits an output signal and self-inhibits via feedback inhibition. The process then repeats itself. (Adapted from Figures 20 and 34 in Grossberg, 1978b).

Accumulating neurobiological evidence supports the interactions predicted by ART, CogEM, and Item and Order working memories. In this view, visual and verbal object categories may be learned in temporal and ventromedial prefrontal (e.g., orbitofrontal) cortex, with the latter responding to combined item and motivational signals, followed by the loading of these item representations into a sequential working memory that codes temporal order information in ventrolateral and/or dorsolateral prefrontal cortex (Barbas, 2000; Bor, Cumming, Scott & Owen, 2003; Browning, Inoue & Hernadi, 2005; Chafee & Goldman-Rakic, 2000; Frey, Kostopoulos & Petrides, 2004; Goldman-Rakic, 1987; Hasegawa, Blitz & Goldberg, 2004; Kostopoulos & Petrides, 2003; Ninokura, Mushiake & Tanji, 2004; Petrides, 1991, 2005; Petrides, Alivisatos & Frey, 2002; Ranganath & Blumenfeld, 2005; Ranganath & D'Esposito, 2005; Ranganath, Johnson & D'Esposito, 2003; Rolls, 2004; Schoenbaum & Roesch, 2005; Schoenbaum & Setlow, 2001).

Given this anatomical interpretation, lateral prefrontal cortex should be relatively insensitive to low-level features, but should show load effects as more items are simultaneously stored, even when there are no learning effects, as various experiments have shown (Habeck, Rakitin, Moeller, Scarmeas, Zarahn, Brown & Stern, 2005; Narayanan, Prabhakaran, Bunge, Christoff, Fine & Gabrieli, 2005; Ranganath, DeGutis & D'Esposito, 2004; Rypma, Berger & D'Esposito, 2002, Song & Jiang, 2005; Woodward, Cairo, Ruff, Takane, Hunter & Ngan, 2005). Even with

similar amounts of information, one can expect to see increased activity of lateral prefrontal cortex if grouping of multiple items on the basis of temporal order is required (Bor, Cumming, Scott & Owen, 2004; Bor, Duncan, Wiseman & Owen, 2003).

Prefrontal cortex is disproportionately large and extensively interconnected in humans (Barbas, 2000). What parts of prefrontal cortex are devoted to storage of cognitive and motor working memories? Model properties are consistent with the hypothesis that non-spatial cognitive working memory in the primate occurs in the ventrolateral prefrontal cortex, with spatial cognitive working memory in the principal sulcus, or the human homolog (Figure 1). Motor working memory may be located in the dorsolateral prefrontal cortex—in particular, outside of and likely dorsal to the principal sulcus in the monkey—and medial premotor cortex—for example, supplementary motor area (SMA) and pre-SMA. Lesion studies support these functional localizations (Curtis & D'Esposito, 2004; Goldman-Rakic, 1987; Passingham, 1993).

These localization hypotheses clarify how bilateral lesions to ventrolateral prefrontal cortex disrupt learned associations of high-level stimulus cues and their planned motor responses (Bussey, Wise & Murray, 2001; Wang, Zhang & Li, 2000). Similar effects occur with lesions that disrupt connections between inferotemporal cortical centers in one hemisphere and ventrolateral prefrontal cortices in the other (Eacott & Gaffan, 1992; Parker & Gaffan, 1998; for a review, see Curtis and D'Esposito, 2004).

Lesions to the lateral prefrontal cortex that include the ventrolateral prefrontal cortex, in addition to the dorsolateral prefrontal cortex, result in an additional deficit in object alternation tasks (Mishkin & Pribram, 1955, 1956; Mishkin, Vest, Waxler & Rosvold, 1969; Pribram & Mishkin, 1956; Petrides, 1995), and selective lesions to areas ventral to the principal sulcus result in impairment on non-spatial working memory tasks, which localized principal sulcal lesions do not impair (Mishkin & Manning, 1978; Passingham, 1975).

Various parts of ventrolateral and dorsolateral prefrontal cortex that are relevant to the processing of serial order are specialized for language understanding and performance. For example, in humans, lesions to Broca's area, which frequently include caudal portions of ventrolateral prefrontal cortex, ventrolateral premotor cortex and anterior insula (Dronkers, 1996; Hillis, Work, Barker, Jacobs, Breese & Maurer, 2004), produce Broca's aphasia. Patients with such lesions have profound articulation difficulty, perhaps due to problems unpacking speech motor codes by ventral premotor and insular cortex. They also typically have difficulty with word naming, produce sentences of two, or sometimes one, words and have dramatically impaired spontaneous speech and repetition competency (Goodglass & Kaplan, 2001).

In contrast, transcortical motor aphasia occurs following lesions slightly superior and anterior to Broca's area, and likely severs its connections with portions of the supplementary motor area, premotor cortex and dorsolateral prefrontal cortex. Such lesions result in difficulty initiating and organizing structured responses, but preserve repetition and cue-based (confrontational) naming (Goodglass & Kaplan, 2001). These data suggest a more specific deficit in sequential organization of verbal motor responses, while sparing the ability to stage low-level repetition or single item confrontational naming. In contrast, Broca's lesions preserve only minimally organized (e.g. one or two word) sequences but notably also remove low-level competency in word naming and repetition.

Model mechanisms may also help to clarify why neuroimaging data show that ventrolateral prefrontal cortex is preferentially activated when a decision needs to be made regarding the relative familiarity of stimuli, whereas dorsolateral prefrontal cortex may be more active in a monitoring condition wherein a non-familiar stimulus must be chosen (Petrides, 2005); see

Section 4. For additional review of these process localizations in brain anatomical regions, see Supplementary Material, Section 2.

2.4. Working Memory Enables Stable List Learning and Memory. LIST PARSE further develops a model of working memory and sequence chunking that was introduced in Grossberg (1978a, 1978b). These articles, together with Grossberg (1976a, 1976b, 1980), laid foundations for several cognitive models, including interactive activation, competitive learning, and competitive decision-making models (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982; Rumelhart & Zipser, 1986; Usher & McClelland, 2001), that receive a laminar cortical interpretation in the LIST PARSE model.

A key issue for the theory of working memories concerns how an Item and Order working memory can carry out short-term storage of event sequences. Grossberg (1978a, 1978b) linked the problem of short-term storage of event sequences to the problem of how the brain carries out *stable learning and long-term memory of list categories*, also called chunks or sequential plans, that are activated by familiar sequences of stored items. This analysis predicts that the same basic circuit design will be used for working memory storage in all modalities, which clarifies why both spatial and non-spatial working memories across modalities seem to share many behavioral and neurobiological properties; see below. This prediction makes precise the common-sense observation that a working memory would be of limited value if it could not support stable learning of sequential plans. If it could not, then every stored sequence, no matter how often it was experienced, would always seem to be novel, and would always be processed as an unfamiliar and non-unitized sequence of items.

The key innovations of the Item and Order model, which is also called the STORE (Sustained Temporal Order REcurrent) model (Bradski, Carpenter, & Grossberg, 1992, 1994), are as follows (see Figure 4):

Relative Activity Codes Temporal Order in Working Memory. Item and Order models predict that a series of events, occurring one after another in time, is encoded in working memory by an *analog spatial pattern of activation that evolves in parallel across a network of content-addressable cells*. This representation represented a radical break from the popular model of Atkinson & Shiffrin (1968, 1971), which proposed binary activations of a series of linearly ordered "slots" wherein each item moves to the next slot as additional items are stored. Temporal order in an Item and Order model is encoded by the relative activity of each content-addressable item, with higher activities being performed earlier (see Figure 4).

Working Memory Design Enables Stable Learning of List Chunks. Given that Item and Order models are part of a larger theory about how working memories are designed to support stable learning and long-term memory of list categories, it remains to explain how these list categories are learned. As noted above, list categories, just like item categories, may be learned as ART categories (see Section 2.1). The list categories are activated by item category sequences that are stored as spatial patterns in a working memory, rather than, as in the case of item categories, by spatial patterns of item features. Multiple subsequences of a stored sequence can bid to learn list categories, or chunks. The brain must therefore solve the *temporal chunking problem*: How can a novel list chunk be learned, under unsupervised learning conditions, from a sequence of stored items whose subsequences have already learned to code their own list chunks? For example, how can a list chunk for the novel word MYSELF be learned even when there may already be strong learned representations for MY, SELF, and/or ELF?

A *masking field* network (see Section 4.1.4) helps to explain how a novel sequence of items in working memory can trigger unsupervised learning of a new list chunk even if some of its

subsequences are already coded by their own list chunks. A masking field's parameters favor longer lists, up to a maximum list length, and can thereby overcome the learned salience of already familiar chunks that code for shorter lists (Cohen & Grossberg, 1986, 1987; Grossberg, 1978a; Grossberg & Myers, 2000). A laminar cortical version of a masking field (layer C_j in Figure 2 and Appendix equation (5)) is proposed within the LIST PARSE model for this purpose. The temporal chunking problem also concerns how these new "superset" list chunks (e.g., MYSELF) can be learned without causing catastrophic forgetting of previously learned "subset" list chunks (e.g., MY and SELF). This issue can be resolved as follows.

LTM Invariance and Normalization Rule. Item and Order models embody two constraints on the design of short-term working memories to ensure that stable bottom-up learning and long-term memory of list chunks can occur. The *LTM Invariance Principle* and the *Normalization Rule* prevent bottom-up learning of superset list chunks from erasing previous memories of subset list chunks. Section 2.5 describes how these two constraints can be realized.

Learned Top-Down Expectations, List Priming, and Fluent Sequential Recall. As in the case of item category learning, every list chunk can learn a top-down expectation through feedback connection weights from the chunk to the items that activated it from working memory (pathway C_j to Y_i in Figure 2 and Appendix equation (3)). This top-down expectation helps to stabilize bottom-up list category learning in the same way that top-down expectations stabilize the learning of item categories. As noted above, the top-down expectation from a list chunk to its item sequence in working memory reads out the spatial pattern of activation that triggered bottom-up learning of the list category. This form of temporal ART dynamics has been used to quantitatively simulate challenging data about speech perception and categorization; e.g., Grossberg, Boardman, & Cohen (1997) and Grossberg & Myers (2000).

When a list chunk is activated either top-down via a volitionally-controlled rehearsal event, or by bottom-up stimuli from another modality (e.g., seeing an object before naming it), its top-down expectation can reactivate in working memory the temporal order of the items that it represents. In this way, unitized list chunks can control fluent rehearsal of familiar item sequences.

Rehearsal. Item and Order models predict that a volitionally-controlled, nonspecific rehearsal wave equally activates all the working memory cells to initiate recall, with the most active cell generating its output first (see V_m in Figure 2 and Appendix equation (18)). This happens whether the working memory is activated bottom-up by sequences of individual items, or top-down by a list chunk.

Rehearsing the most active item first immediately raises the problem of *perseveration*; namely, what keeps the most active item from being rehearsed over and over again, thereby preventing less active items from being rehearsed? Perseveration can, in fact, occur in some clinical syndromes.

Inhibition of Return. Item and Order models predict that this perseveration problem is solved by *output-contingent self-inhibition*. In other words, each rehearsed cell's output activates a recurrent inhibitory interneuron that shuts off the cell, and thereby enables less active items to be rehearsed in the order of their relative activation (Figure 4 and Appendix equations (9) and (14)). Self-inhibition of each rehearsed item causes *inhibition of return* to that item.

Recurrent Normalization. The activity level of a stored item influences the reaction time with which it is rehearsed, with the largest activity rehearsed first. As items with the largest activities are rehearsed, the remaining items have progressively smaller activities. Item and Order models predict how the stored pattern can renormalize its total activity across all the active

cells through time as items are rehearsed. Such activity renormalization at least partially compensates for the slow-down in reaction time due to small activities. This Normalization Rule is realized by a recurrent, or feedback, network, whose properties will be further defined below. Indeed, such a recurrent network also realizes the main property of working memory, namely persistent storage of item sequences through time.

Variants of the above hypotheses have found their way into many ordinal modeling studies of working memory and sequential performance (e.g., Boardman & Bullock, 1991; Houghton & Hartley, 1996; Page & Norris, 1998; Rhodes et al., 2004).

2.5. Stable Learning Implies Primacy, Recency, and Bowed Activation Gradients. As noted above, Item and Order working memories may be derived from the LTM Invariance Principle and the Normalization Rule. The LTM Invariance Principle is the main postulate to ensure that a new superset list chunk can be learned without forcing catastrophic forgetting of familiar subset list chunks. As a result, subset list chunks can continue to activate their familiar list chunks until they are inhibited by contextually more predictive superset list chunks; e.g., until MY is supplanted by competition from MYSELF through time. The *learning* of chunk MY within its bottom-up filter is not undermined, but the *current activation* of the chunk MY can be inhibited by MYSELF. Mathematically, this boils down to the following property: activities of items in working memory preserve their *relative* activations, or *ratios*, throughout the time that they are stored in working memory.

The Normalization Rule assumes that the *total activity* of the working memory network has a maximum that is (approximately) independent of the total number of actively stored items. In other words, working memory has a *limited capacity* and activity is redistributed, not just added, when new items are stored.

Together, these simple rules generate working memories that can support stable learning and long-term memory of list chunks. It was mathematically proved that, under constant attentional conditions, the pattern of activation that evolves in an Item and Order working memory is one of following types (Bradski, Carpenter and Grossberg, 1992, 1994; Grossberg, 1978a, 1978b):

P*rimacy Gradient.* Here, the first item to be stored has the largest activity and the last item to be stored has the smallest activity. A primacy gradient allows the stored items to be rehearsed in their presented order.

R*ecency Gradient.* Here the first item is stored with the smallest activity and the last item with the largest activity. Rehearsal of a recency gradient recalls the most recent item first and the first item last.

B*owed Gradient.* Here, the first and last items to be stored have larger activities, and thus are earlier rehearsed, than items in the middle of the list.

From Primacy to Bowed Gradient. It was also proved that, as more and more items are stored, a primacy gradient becomes a bowed pattern whose recency part becomes increasingly dominant.

This last result predicted a totally unexpected reason for the ubiquitous occurrence of bowed gradients in many types of serially ordered behavior: The property of stable learning and memory of list chunks imposes a severe limitation on the number of items that can be recalled in the correct temporal order from working memory, due to the development of a bow in the stored gradient of sufficiently long lists. The ability of LIST PARSE to develop recency gradients in response to sufficiently long lists helps to explain many data about free recall.

All of these postulates have received support from subsequent psychological and neurobiological experiments. These data will be discussed in greater detail below. It is worth

noting now the conclusion that Farrell & Lewandowsky (2004) derived from their data: "Several competing theories of short-term memory can explain serial recall performance at a quantitative level. However, most theories to date have not been applied to the accompanying pattern of response latencies…these data rule out three of the four representational mechanisms. The data support the notion that serial order is represented by a primacy gradient that is accompanied by suppression of recalled items". Jones, Farrand, Stuart, & Morris (1995) reported similar performance characteristics to those of verbal working memory for a spatial serial recall task, in which visual locations were remembered. Agam, Bullock, & Sekuler (2005) reported psychophysical evidence of Item and Order working memory properties in humans as they perform sequential copying movements, thereby supporting the prediction that all working memories obey similar laws. Agam, Galperin, Gold, & Sekuler (2006) reported data consistent with the formation of list chunks as movement sequences are practiced, thereby supporting the prediction that working memory networks are designed to interact closely with list chunking networks. Averbeck, Chafee, Crowe, & Georgopoulos (2002, 2003a, 2003b) reported the first neurophysiological evidence in monkeys that a primacy gradient, together with inhibition of the most active cell after its command is read out, governs the sequential performance of sequential copying movements. The fact that verbal, spatial, and motor sequences, in both humans and monkeys, seem to obey the same working memory laws provides accumulating evidence for the Grossberg (1978a, 1978b) prediction that all working memories have a similar design to enable stable list chunks to form.

2.6. Stable Learning Implies the Magical Numbers Four and Seven. The prediction that primacy gradients become bows for longer lists provides a conceptually satisfying explanation of the well-known *immediate memory span* of 7 +/- 2 items (Miller, 1956). It was originally used to explain data about free recall and related paradigms in which bowing effects are observed (Grossberg, 1978b). Indeed, because relative activity translates into both *relative order* and *probability of recall* (bigger activities can provide more reliable recall in a noisy brain), such a model helps to explain why items from the beginning and end of a list in free recall may be recalled earlier and with larger probability (Murdock, 1962). *Transposition errors* also have a natural explanation in such a working memory, since stored items with similar activity levels will transpose their relative activities, and thus their rehearsal order, more easily than items with very different activity levels if noise perturbs these levels through time. Grossberg (1978a, 1978b) also proved that, if attention varies across items, then multi-modal bows, or Von Restorff (1933) effects, also called isolation effects (Hunt & Lamb, 2001), can be obtained by altering the relative sizes of stored activities. Von Restorff effects can also be caused by rate and feature similarity differences across items, factors which also influence bowing in the present modeling framework. Associative and competitive mechanisms that are consistent with the LIST PARSE model have been used to explain Von Restorff effects during serial verbal learning (Grossberg, 1969, 1974).

The Grossberg (1978a) analysis distinguished between the classical *immediate memory span* (IMS) of Miller (1956) and the then new concept of the *transient memory span* (TMS). The TMS was predicted to be the result of short-term working memory storage recall, without a significant top-down long-term memory (LTM) component. The TMS is, accordingly, the longest list length for which a working memory can store a primacy gradient. The IMS was predicted to be the result of combining bottom-up inputs and top-down read-out of list chunk learned expectations on the relative activities stored in working memory, and thus the temporal order that is recalled (see Section 2.4). The read-out of list chunk top-down expectations into working

memory was predicted to generate an extended primacy gradient in working memory, and thus a longer list that could be recalled in the correct order. The role of read-out from long-term memory into short-term memory plays a significant role in current dual-store models; see Section 7.1.

The extended list chunk primacy gradient due to top-down read-out of list chunk expectations results from several factors. First, bottom-up inputs for sufficiently short lists tend to store a primacy gradient in short-term working memory. The list chunks that are activated at the beginning of a list thus tend to learn a primacy gradient in LTM. Moreover, these early-activated list chunks and their primacy gradient in short-term memory are active for a longer time than later list items, which enables their learning to be stronger, and thus further favors primacy dominance. Top-down read-out of these primacy gradients from LTM combines with bottom-up short-term working memory primacy gradients to cause longer primacy portions of the stored list, thereby leading to the veridical serial recall of more items, and thus a longer observed memory span.

Based on such considerations, Grossberg (1978a) proved that the TMS is smaller than the IMS. Estimating the IMS at seven, he predicted that the TMS would be around four. Cowan (2001) has reviewed experimental data that support the existence of a four plus-or-minus one working memory capacity limit when LTM and grouping influences are minimized, which is consistent with this prediction. Indeed, long-term memory (LTM) does bias working memory toward more primacy dominance (e.g. Knoedler, 1999), and its influence can be difficult to limit. Cowan (2001) reviewed proposals for limiting LTM influence, such as using novel sequence orderings of well-learned items that are difficult to group.

2.7. Working Memories from Recurrent On-Center Off-Surround Networks. Item and Order models provide a satisfying answer to the following perplexing question: How could evolution have been smart enough to discover a brain design for something that seems as sophisticated as a working memory? Grossberg (1978a, 1978b) proposed how a ubiquitous and ancient brain circuit design can be specialized to realize the properties that are needed in an Item and Order working memory; namely, a *recurrent on-center off-surround network* of interactions among cells that obey the *membrane, or shunting, equations* that neurons are known to obey (Grossberg, 1973, 1980; Hodgkin & Huxley, 1952; see equations (1), (2) and (3) in the Appendix).

In such a model, feedback, or *recurrent*, connections help to store inputs in short-term memory after the inputs shut off. These feedback signals are of two types: Self-excitatory, or *on-center*, signals from each cell to itself attempt to maintain the cell's activation after an external input shuts off (see Figure 2 feedback pathways between X_i and Y_i). Competitive, or *off-surround*, lateral inhibitory signals to other cells in the network balance the self-excitatory signals (see Figure 2 pathways from Y_i to X_k with $k \neq i$). On-center off-surround networks of cells that obey shunting dynamics solve a ubiquitous problem that is faced by all neuronal networks, both in the brain and all other biological tissues: the *noise-saturation dilemma* (Grossberg, 1973, 1980). Namely, such networks keep their stored activities large enough to avoid being distorted by internal cellular noise, yet not so large as to activate cells maximally, saturate their responses, and destroy a record of analog input differences. A shunting on-center off-surround network can contrast-normalize and preserve the analog sensitivity of its cell activities, no matter how large the inputs to the network are chosen.

Remarkably, a shunting on-center off–surround network that is designed solve the noise-saturation dilemma can, if it has *recurrent* on-center off-surround connections, also store a

sequence in working memory in a manner that obeys the LTM Invariance Principle and the Normalization Rule. This is due to the way that shunting, or multiplicative, interactions compute *ratios* of cell activities across the network, as is required by the LTM Invariance Principle. Such a recurrent network behaves like an Item and Order working memory model when it is equipped with a volitionally-activated nonspecific rehearsal wave to initiate read-out of stored activity patterns, and output-contingent self-inhibitory feedback interneurons to prevent perseverative performance of the most activity stored item (Figures 3 and 4). Bradski, Carpenter & Grossberg (1992, 1994) mathematically prove these properties of Item and Order working memory storage in the STORE working memory model, which defines a particularly simple shunting recurrent on-center off-surround network that enables complete theorems to be proved.

2.8. Related Modeling Studies. Since the 1970's, models of working memory have undergone significant development. The seminal work of Baddeley & Hitch (1974) and Baddeley (1986) on the phonological loop in working memory has had a particularly large impact. Houghton (1990) introduced the term Competitive Queuing, or CQ, to refer to models that utilize the Item and Order concept of storing and rehearsing queues of items in working memory using competitive interactions based upon the relative strengths of an encoded item attribute. Given that all sequential working memories need to be able to support stable learning and memory of sequence chunks or plans, it is perhaps not surprising that models utilizing Item and Order working memory storage concepts have been used to explain and simulate several different types of learning data, including the learning of musical sequences (Mannes, 1994; Page, 1994), speech perception and word recognition (Cohen & Grossberg, 1986, 1987; Grossberg, 1978b, 1986; Grossberg, Boardman & Cohen, 1997; Grossberg & Myers, 2000; Grossberg & Stone, 1986b), attention shifting during working memory storage (Grossberg & Stone, 1986a), control of cursive handwriting (Bullock, Grossberg & Mannes, 1993), and reaction time data during skilled typing (Boardman & Bullock, 1991; Rhodes et al., 2004).

A particularly relevant model was described by Page & Norris (1998). These authors used a primacy gradient in working memory combined with competitive queuing, or inhibition of return, to explain and simulate cognitive data about immediate serial order working memory, such as word and list length, phonological similarity, and forward and backward recall effects. However, this study *assumed* a primacy gradient, and did not attempt to model the network dynamics that generate it as an emergent property. The Farrell and Lewandowsky (2004) study, mentioned in Section 2.5, examined the latency of responses following serial performance errors. They asserted that (p. 115): "The data support the notion that serial order is represented by a primacy gradient that is accompanied by suppression of recalled items."

Despite a long-term interest in working memory tasks (Nipher, 1878), the first conceptual models arose relatively recently (Atkinson & Shiffrin, 1968, 1971; Baddeley & Hitch, 1974; Estes, 1972; Glanzer, 1972; Glanzer & Cunitz, 1966; Grossberg, 1978a, 1978b; Lee & Estes, 1977; Postman & Phillips, 1965). Detailed mathematical models attempting to account for large cross-sections of the known properties of a particular working memory task have been even more recent (e.g., Anderson & Matessa, 1997; Brown, Preece & Hulme, 2000; Burgess & Hitch, 1999; Farrell & Lewandowsky, 2002; Henson, 1998b; Howard & Kahana, 2002; Lewandowsky, 1999; Lewandowsky & Murdock, 1989; Nairne, Neath, Serra & Byun, 1997; Page & Norris, 1998; Raaijmakers & Shiffrin, 1980; Ratcliff, 1978, 1981; Rhodes et al., 2004). These models were developed to capture detailed aspects of either serial or free recall literatures. LIST PARSE bridges these separate literatures and uses design principles that may be interpreted in terms of known brain processes.

2.9. Human Cognitive and Monkey Sensory-Motor Data as Emergent Properties of Laminar Prefrontal Cortical Circuits. The LIST PARSE model joins together a cognitive working memory, a motor working memory, a motor trajectory generator, and a volitional controller that coordinates variable-rate recall of both novel and previously learned sequences as they are read-out into sequential action from a cognitive plan (Figures 1 and 2).

These circuits are used below to first describe quantitative simulations of human cognitive data about serial recall and immediate free recall using the same model parameters. Then the model is used to simulate the neurophysiological data of Averbeck et al. (2002, 2003a, 2003b; also see Rhodes et al., 2004) from the peri-principalis region of dorsolateral prefrontal cortex in macaque monkeys during a sequential copying task. LIST PARSE is thus the first model that links dynamically evolving emergent properties of identified prefrontal laminar cortical circuits with verbal cognitive data from humans about both serial and free recall, and sensory-motor data from monkeys.

3. Experimental Background. This section reviews key experimental data that LIST PARSE can explain, and highlights several current debates over the explanation of these data in the literature. In this section's parts below, we (1) introduce the modeled protocols, (2-4) discuss several reasons why it is difficult to compare data about serial recall and immediate free recall; (5-7) describe the simulated serial recall data; (8) highlight the debate between interference-based and decay-based accounts of working memory limits, centered largely around the word-length effect in serial recall; (9) describe the simulated immediate free recall data; (10) discuss performance differences between serial and free recall; (11) discuss evidence for multiple memory stores in free recall (e.g., dual-store accounts; single store accounts are discussed in Section 7); (12) discuss data concerning the influence of distracters; (13) discuss the similarity of spatial serial recall to the non-spatial serial recall data that was simulated and, finally (14) introduce the sequential movement task that was simulated.

3.1 Immediate Serial Recall and Immediate Free Recall Experiments. The two most extensive literatures to be addressed by an integrated model of working memory function are those of the immediate serial recall (ISR) task (e.g., Henson, Norris, Page & Baddeley, 1996) and the immediate free recall (IFR) task (e.g., Tan & Ward, 2000). In the ISR protocol, a list of items (typically words, letters, numbers or spatial targets) is presented (either auditory or visual) in succession, with identical stimulus durations and inter-stimulus intervals followed by a brief retention interval. Participants are then cued to recall the items in the same serial order in which they were presented (see Figure 5). In the IFR protocol, again such a list of items is presented. However, participants are cued to recall them in any order they choose (see Figure 5). A number of significant complications hamper analysis and comparison of the ISR and IFR task literatures. Among these are: differences in the definition of errors, difficulty in isolating long-term memory influences and chunking, and difficulty in monitoring or preventing rehearsal.

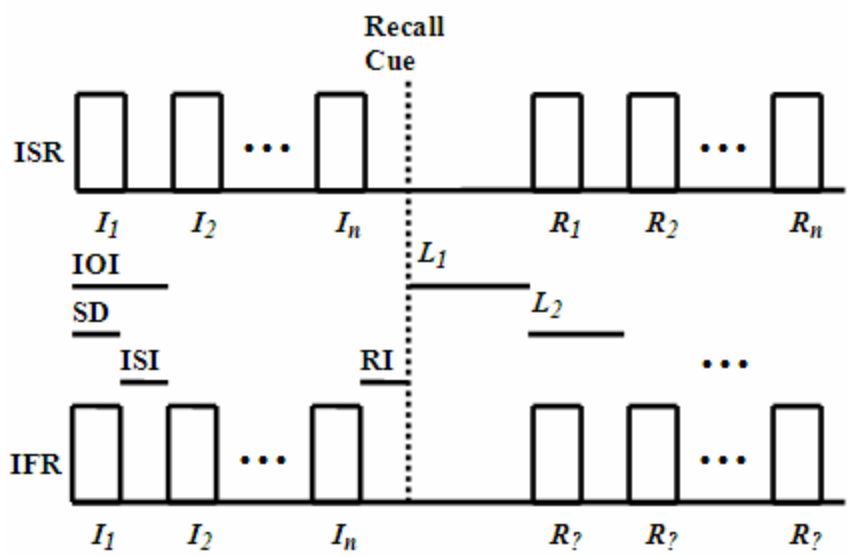


Figure 5. The immediate serial recall (ISR) and immediate free recall (IFR) protocols both consist of a series of n item inputs (I_1, …, I_n), typically with the same stimulus duration (SD) and inter-stimulus interval (ISI). The inter-onset interval (IOI) equals the sum of SD and ISI. A brief retention interval (RI) occurs before a cue is presented to begin recall. In the ISR protocol, participants make responses (R_1, …, R_n) corresponding to stimuli in their presented order. In the IFR protocol, participants make responses corresponding to the stimuli in any order they choose. Subject responses occur with variable latency times (L_1, L_2, …, L_n).

3.2. Different Span and Total Error Definitions. The main difficulty in comparing the ISR and IFR tasks is that different experimental protocols are used. For instance, *span length* for a given inter-stimulus interval (ISI), stimulus duration (SD), inter-onset interval (IOI)—the sum of SD and ISI—and retention interval (RI) (see Figure 5) refers to the list length at which a participant makes some percentage of list recall errors. It is typically the length of list studied. However, the methods of computing this span, even within a protocol (e.g., the ISR task), are experimenter-specific, ranging from the maximum list length at which a subject can perform two consecutive lists correctly (Cowan et al., 1999) to variations upon the number of items that yield approximately 50% of lists recalled correctly (Hulme, 1991). This leads to different numbers of items and levels of performance being taken to represent "span." Such differences hamper comparisons between studies even within the same protocol.

Different span definitions across protocols are further complicated by the different kinds of error in each protocol. In IFR, item errors are the only error types, including *intrusions* of items that were not in the presented list, and *omissions* of items that were presented. Intrusions and omissions occur in ISR too, but *order errors* also occur in which an item is successfully recalled, but in the wrong serial position, or is erroneously repeated. IFR and ISR studies commonly report percentage of total correct responses, or total errors, versus serial position, thereby averaging across the different possible error types within each protocol.

Henson, Norris, Page & Baddeley (1996) did perform an analysis of the types of ISR errors. Most errors are order errors, primarily transpositions, for lists of span length (or shorter). Thus, the majority of ISR errors at span are not errors in an IFR task. Using total list error rate (span) to

determine experimental parameters, even with identical ISR and IFR protocols, is complicated by this fact.

3.3. Chunking and Long-term Memory Read-out. It is also difficult to factor out long-term memory and chunking influences, whether volitionally controlled or otherwise. A common presumption has been that, during *immediate* recall protocols, these effects are minimal. However, Cowan (2001) summarized data showing that these effects are common, and have masked a four plus-or-minus one chunk capacity underlying purely short-term memory effects; cf., the Transient Memory Span in Section 2.6.

Cowan (2001) suggested that the "magical number seven plus-or-minus two" of Miller (1956) was an artifact of experimental protocols that allow easy learned unitization of items (e.g., digits with which subjects have extensive experience) into list chunks, thereby allowing the extension of the four plus-or-minus one limit to approximately seven items. LIST PARSE builds upon Item and Order model analyses of why such chunk capacity limits exist, and are necessary for stable learning and long-term memory of temporal order information (Section 2.5).

In support of this approach, recent studies provide evidence that list chunks code both item and order information. In Cowan, Chen & Rouder (2004), associative pairs were presented with varying frequencies across trials (0-4 exposures) within the context of an 8-item ISR task. Subjects recalled more items when a test list contained some of the previously seen paired-associates, and performance scaled with increasing amounts of prior exposure. Moreover, subjects showed nearly identical item and total (i.e., item plus order) error improvement with increasing exposure. That is, prior exposure to paired items resulted in subjects remembering more items in lists that included them, but those additional recalled items did not result in any additional order errors. The sequential order of the paired items was thus intrinsic to their learning. Subjects may also process statistical item-position associations. That is, an item may be statistically better learned if it repeatedly in the same absolute list position, regardless of the ordering of other items in the list (Conway & Christiansen, 2001).

3.4. Controlling Rehearsal Strategy. It is important to control, monitor, or prevent, covert or overt rehearsal during list presentation. Four common protocols are: (1) overt rehearsal in which subjects are instructed to vocally rehearse between stimulus presentations; (2) articulatory suppression protocols in which subjects are instructed to mouth, or vocalize, some repetitive phrase (such as "the") when stimuli are presented and between pauses; (3) rapid list presentation protocols which limit the time that covert rehearsal can occur; and (4) distracter protocols which require subjects to perform a concurrent task during the retention interval, throughout the task, or both. How different rehearsal strategies and recall requirements of ISR and IFR tasks interact with a shared LIST PARSE working memory circuit are used to explain similarities and differences between the two tasks.

3.5. Bowing of Serial Position Performance Curves. Bowing of serial position error curves was observed over a century ago (Nipher, 1878) and needs to be explained in any model of working memory. Both ISR and IFR exhibit bowed error curves with a *primacy* portion at the beginning of the list, in which errors increase with item position, and a *recency* portion at the end of the list, in which errors decrease with serial position.

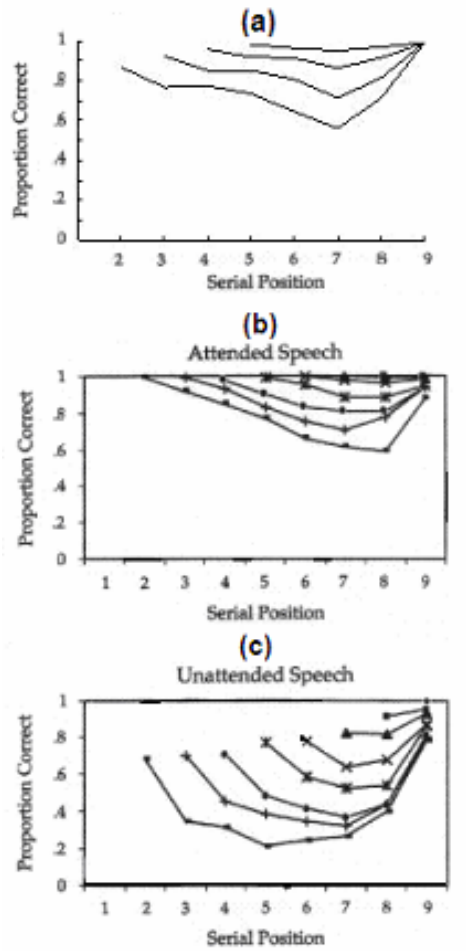


Figure 6. (a) Simulations of proportion correct vs. serial position by the LIST PARSE model for auditory presentation of various list lengths. (b) Observed mean proportion of items correct in attended speech (auditory presentation) ISR task vs. serial position at various list lengths. (c) Observed mean proportion of items correct in unattended speech (auditory presentation) ISR task vs. serial position at various list lengths. Note: all graphs are shifted rightward to avoid overlaps. (Data reprinted with permission from Cowan et al., 1999).

The typical ISR serial position curve at approximately span list length (e.g., often defined as the length resulting in approximately 50% of lists being recalled entirely correctly) consists of an extended primacy portion, with approximately linear increases in total errors with item position, and a smaller recency portion. With auditory presentation, the recency portion may extend to several items at the list end, depending upon the experimental design (Figures 6b and 6c). With visual list presentation, the recency portion is often just one item long (Figure 7, circles). However, these one, or two, item error patterns are not universal (Drewnowski, 1980; Drewnowski & Murdock, 1980), as some modelers (Page & Norris, 1998) have suggested, and their shape can be altered by experimental variables like attentional modulation (Figure 6c). This fact calls into question the sufficiency of recency explanations that cannot incorporate effects of attention (e.g., Page & Norris, 1998).

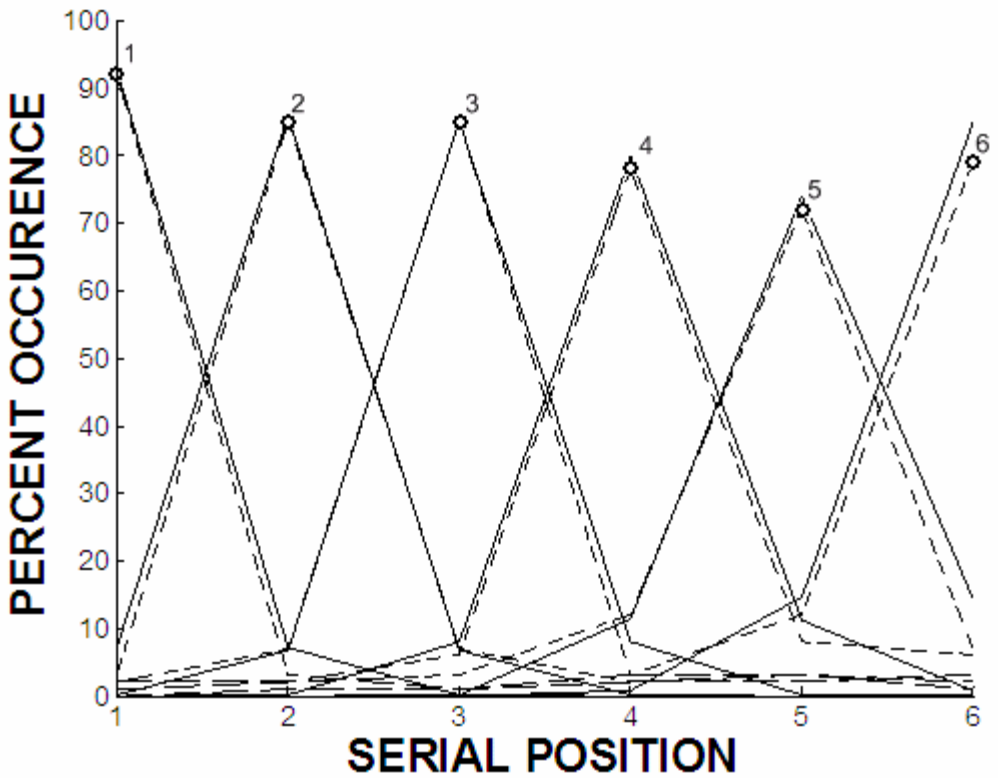


Figure 7. Percentage of recalls at each item *recall* serial position vs. item *presentation* serial position. Solid Lines: Simulations by LIST PARSE model. Dashed lines: Data from Experiment 1 in Henson et al., 1996. Circles: Percentage of correct responses for each serial presentation position. (Data reprinted with permission from Henson et al., 1996).

3.6. Error Type Distributions: What Causes Recency Gradients? Few ISR analyses (e.g., Drewnowski, 1980) study the different types of error comprising the serial position curve. Henson et al. (1996) did so and showed that: (1) for sub-span and span length lists, omissions monotonically increased with serial position; (2) order errors comprised the vast majority of errors for lists approximately at span; (3) transposition errors (i.e., swapping of item serial positions) caused the majority of order errors; (4) transpositions of neighboring items were far more likely than between remote positions (Figure 7); and (5) when an item is not recalled in its correct serial position, the displaced item is by far the most likely item to be recalled in the next position (three times more likely than others), a property that Norris, Page & Baddeley (1994) termed *filling-in* (not to be confused with lightness or color filling-in in the context of visual perception). Page & Norris (1998) noted that, if one assumes a primacy gradient, then edge effects can explain decreased order error rates for the first and last items and explain the observed short recency portions in span length ISR. This happens because the first and last item in a primacy gradient have only one similarly active neighboring item with which to interchange, and thus are less susceptible to transposition errors than other list items.

Monotonic increase in *item* errors with serial position does not occur for super-span lists with auditory presentation (Klein, Addis & Kahana, 2005) or visual presentation (Drewnowski, 1980; Drewnowski & Murdock, 1980; Watkins & Watkins, 1980). Here, extended recency gradients in *item* errors may occur. The auditory recency gradient could perhaps be explained by a

specialized echoic memory. However, such an explanation cannot explain the visual recency gradient.

This recency gradient in *item* errors when list length exceeds span is not explained by other models of serial order, which assume item errors to be monotonically increasing with serial position, or flat, with one known recent exception (Henson, 1998b). Such a bowed pattern of item errors was mathematically proved to occur in Grossberg (1978a, 1978b) and Bradski, Carpenter, and Grossberg (1992, 1994) whenever a sufficient number of items are stored by a working memory that obeys the LTM Invariance Principle and the Normalization Rule (see Section 2.5). Such a bowed gradient is not postulated. Rather, it is an emergent property of the recurrent shunting on-center off-surround network dynamics that realize the working memory (see Section 2.7). Bowed working memory gradients for super-span lists also emerge from the laminar ventrolateral prefrontal cortical circuits that realize LIST PARSE cognitive working memory.

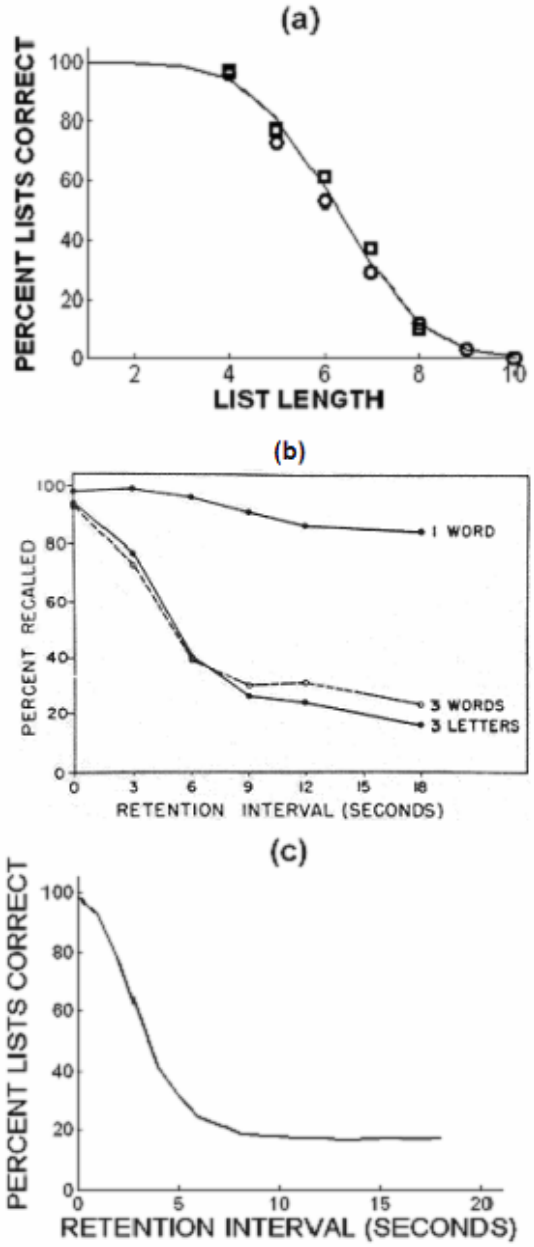


Figure 8. (a) Percentage of lists recalled correctly during immediate serial recall tasks for various list lengths. Solid curve: Simulation of LIST PARSE model for lists of varying numbers of items. Circles: data averaged over limited and unlimited letter conditions from Crannell & Parrish, 1957. Squares: data reprinted with permission from Baddeley & Hitch, 1975. (b) Percentage of lists recalled correctly after extended distracter-filled retention intervals (data from Murdock, 1961). (c) Simulations of LIST PARSE model for 3 item lists with various filled (no rehearsal) retention intervals. (Reproduced with permission from Baddeley & Hitch, 1975, Crannell & Parrish, 1957, and Murdock, 1961).

3.7. List Length Effects. Other things being equal, lists with more items are more poorly recalled (Figure 8a). Similar deficits have been proposed for phonological lists with items that take longer to articulate (Baddeley, 1986), although this proposal has been recently challenged (see below). Baddeley and colleagues (Baddeley et al., 1975; Baddeley, 1986) have suggested that the maximum list length of phonological material that can be correctly recalled (list span) in the ISR task is limited to the number of items that can be pronounced in approximately 1.5 - 2 seconds. The maximum list length thus scales linearly with articulation rate. These findings have been extended to a variety of item types and across languages (e.g., Schweickert & Boruff, 1986; Zhang & Simon, 1985).

3.8. Temporal Limitations Upon Recall: Decay or Interference? Delays in recall of as little as 2 - 3 seconds can render portions of unrehearsed lists of items at span length irrecoverable (e.g., Baddeley, 1986; Baddeley & Scott, 1971; Baddeley, Thompson & Hitch, 1975; Cowan, Day, Saults, Keller, Johnson & Flores, 1992b). Even for lists of sub-span length, in a serial recall task, a retention interval filled with a distracting task that prevents rehearsal can greatly reduce recall accuracy (Figure 8b).

The facts that span seems to be related to output time and/or articulation rate, and that unrehearsed delays hinder list recall, only imply that recall performance is to some extent *time-dependent*. It does not necessarily mean that a simple *temporal decay* process causes these phenomena. In fact, the explanations given for limitations in working memory performance are nearly as numerous as the experimental protocols used to explore them. Cowan (2001) proposed that at least seven different approaches to explaining the limited capacity of working memory exist, with the largest dichotomy concerning whether time (especially *temporal decay*) or the amount and type of information being simultaneously stored (*interference*) principally limits accurate recall capability. For temporal decay, the ability to recall items declines due to time held in memory prior to item recall. For interference, distracters or other items in the list decrease the ability to recall an item correctly. Various proposals have been made about possible mechanisms (e.g., Cowan, 2001; Jensen & Lisman, 1996; Lewandowsky, 1999; Lewandowsky & Murdock, 1989; Nairne, 1990; Neath & Nairne, 1995; Schweickert & Boruff, 1986; Service, 1998, 2000).

Several studies support temporal decay as a key reason for recall performance limits. By carefully monitoring the timing of items throughout list presentation and recall, their analyses show that the time (*articulation time*: Cowan et al., 1992; or *total output time*: Dosher & Ma, 1998) taken to recall an item from the beginning of the recall period strongly correlates with decreased recall performance. Cowan et al. (1992) found, in particular, that performance declined when words that took longer to articulate were located early vs. late in the list. This conclusion is consistent with the literature about articulation time (Baddeley, 1986; Baddeley & Hitch, 1974; Jarrold, Hewes & Baddeley, 2000; Schweickert & Boruff, 1986). Similarly, Henry

(1991) and Avons, Wright & Pammer (1994) proposed that diminished word length effects in probed recall versus serial recall reflected the shorter output time necessary to complete the task.

Cowan et al. (1992b) studied the responses of participants with differing spans in Cowan (1992a), and found that larger item span correlated with larger temporal extent of inter-response intervals. Cowan concluded that, rather than a single rate of decay occurring during the recall period (i.e., total output time dependence), there may be multiple rates of decay (during articulation) and reactivation (during inter-response pauses) given that subjects who performed better tended to take longer between their articulations. The authors called this a "Decay-and-Reactivate" process, but did not propose mechanisms whereby it may be achieved.

Such studies have not ended the debate in favor of pure temporal decay, particularly articulation time, accounts. Service (1998, 2000) and Lovatt, Avons & Masterson (2000, 2002) have instead asserted the importance of phonological complexity and error recovery to explain how performance declines with total output time (Dosher & Ma, 1998), rather than articulation time. Lovatt et al. (2000, 2002) used different sets of di-syllabic words than were used in previous studies that showed a correlation between performance and articulation time (e.g., Baddeley et al., 1975; Caplan, Rochon & Waters, 1992; Cowan, 1992). They also controlled for the numbers of errors in the first half of list recall. No effect of articulation time (word length) was observed when lists were controlled for error rates. Thus, the previously reported effect of articulation time (as opposed to purely output time) upon recall performance could be an artifact of the particular data set, in which words with longer articulation time had a higher propensity for errors.

In another study skeptical of articulation time results, Bireta, Neath & Suprenant (2006) found no evidence for a general word length effect in lists of mixed word length during tests over a variety of stimulus sets. Although the Cowan (1992) results were replicated, the authors suggested that this word length effect may be specific to that stimulus set, since it was not replicated with five other stimulus sets. Bireta et al. suggested that the long words were less "distinct" in the Cowan data, perhaps due to "their reduced imaginability". Modeling of these different results has been recently presented (for review, see: Hulme, Suprenant, Bireta, Stuart & Neath, 2004).

The Waugh & Norman (1965) classical demonstration of interference effects used a probed serial recall protocol to show that the number of items in a list is negatively related to probability of probed recall, with temporal factors playing only a minor role. Cowan, Wood, Nugent & Treisman (1997b) found dissociable effects of duration (implying temporal effects) and complexity (implicating interference effects) using lists of variable word length in an ISR task. More pointedly, the Baddeley et al. (1975) articulation rate vs. list span effect itself points to the importance of interference, since it establishes a specific maximal capacity (e.g., number of items) that can be correctly recalled for a given articulation rate, rather than establishing an amount of time that an arbitrary amount of information can be maintained in memory (Schweikert & Boruff, 1986).

In fact, when sub-span lists are used, participants can retain this information across filled retention intervals, which presumably prevent rehearsal, for times that far exceed the few second (span length) interval (Figure 8b; Murdock, 1961). For sub-span lists, a positive asymptote in performance level can result over extended observed intervals (Figure 8b). These findings seem inconsistent with continuous temporal decay of an unrehearsed item. The failures to consistently record and report order of recall in the IFR task and to time participant responses during the recall phase of both tasks have hindered resolution of this debate across paradigms.

LIST PARSE provides a detailed neural explanation of these effects (Figure 8). Both interference and temporal decay effects are emergent properties of shunting recurrent on-center off-surround interactions between layers 6 and 4 of the model's working memory in ventrolateral prefrontal cortex (Figure 2). Interference occurs because the presence of other items in working memory competitively limits the amount of activity devoted to each stored item (see Section 2.4). Temporal decay effects occur because representations of items converge through time if no rehearsals occur, thereby reducing their distinctiveness and increasing the probability of order errors (see Figure 12a below).

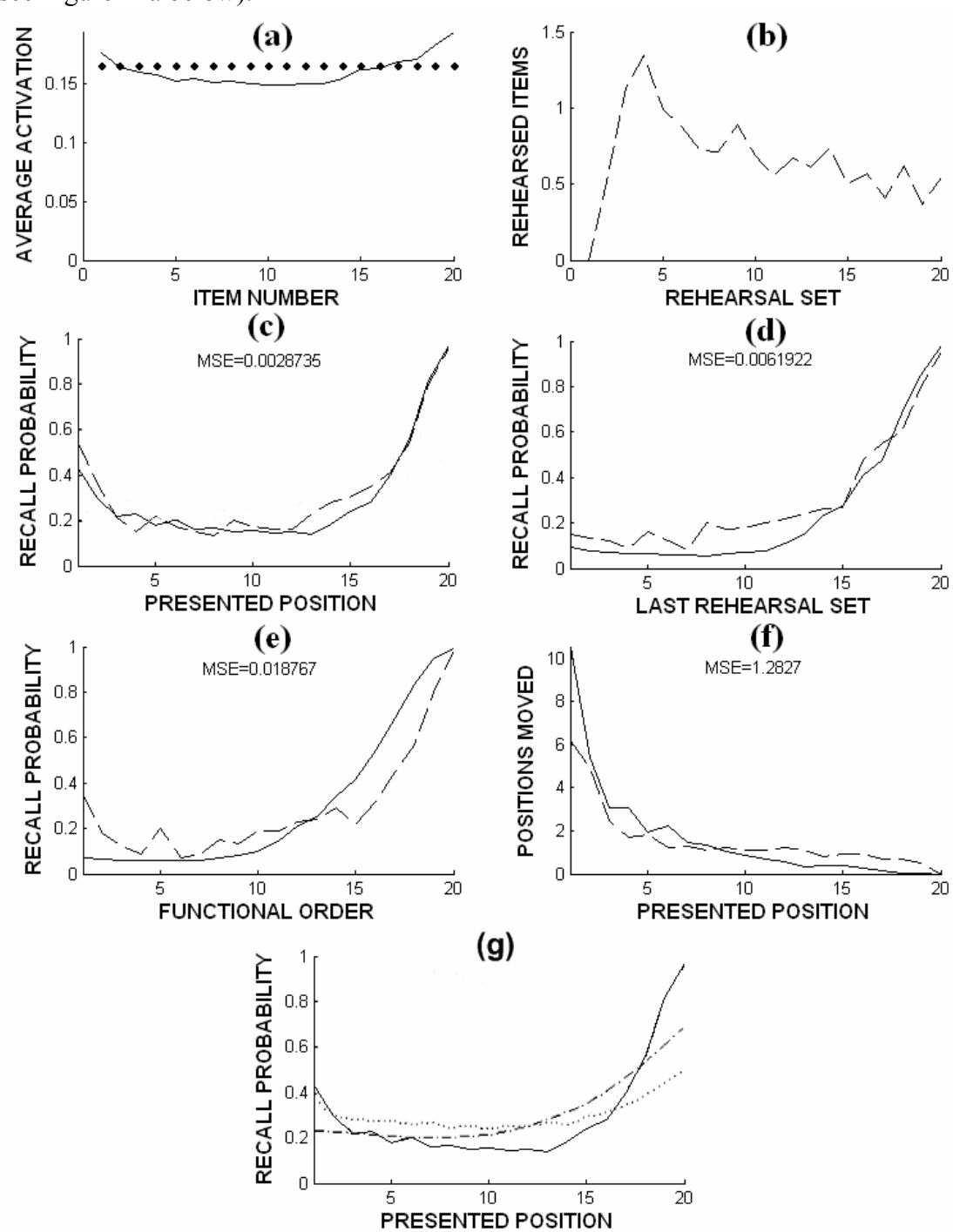


Figure 9. Solid curves in (a) – (f): Simulations of performance measures during an IFR task by the LIST PARSE model. Dashed curves in (a) – (f): Data from the fast condition of Experiment 1 in Tan & Ward, 2000. (a) Average activation of Y_i cells representing each item when errors were assessed and omission threshold (dotted curve). (b) Observed number of rehearsals of distinct previous items during each rehearsal interval in Tan & Ward (2000). (c) Probability of recall vs. presented (nominal) serial position; that is, the standard serial position performance curve. Dotted curve: model prediction of serial position performance curve during delayed free recall; a 12.5 second distracter-task filled interval prevents rehearsal after list presentation before the recall cue. Dashed-dotted curve: model prediction of serial position performance curve during continuous-distracter free recall; rehearsal is prevented both during a 12.5 second post-list distracter period, and between list item presentations. (d) Probability of recall vs. set in which items were last vocalized. (e) Percentage of recall vs. (functional) order of last rehearsal. (f) Difference between set in which an item was last vocalized (rehearsed) and its presentation set. (Data reprinted with permission from Tan & Ward, 2000). (g) Comparison of immediate free recall, delayed free recall, and continuous-distracter free recall. Solid curve: simulated probability of recall in immediate free recall, same as in (c). Dotted curve: model prediction of serial position performance curve during delayed free recall; a 12.5 second distracter-task filled interval prevents rehearsal after list presentation but before the recall cue. Dashed-dotted curve: model prediction of serial position performance curve during continuous-distracter free recall; rehearsal is prevented both during a 12.5 second post-list distracter period and between item presentations.

3.9. Immediate Free Recall Serial Position Curves. Figures 9a – 9f show the data from an IFR task of Tan & Ward (2000, Experiment 1, fast condition) and LIST PARSE simulations of these data. In this task, words were visually presented every 1.25 seconds and subjects were instructed to read the word aloud and to rehearse previously presented words overtly while being tape recorded. Subjects were then required to write down as many words as they could recall for one minute following a recall cue. Model properties are derived entirely from properties of its cognitive working memory in ventrolateral prefrontal cortex (Figures 1 and 2), with no effects of a "transient sensory memory" (see Section 3.10). This is of particular interest in the model's simulations of recency gradients, which in long lists can be 6 or 7 items long, much longer than the capacity of any hypothesized short-term store.

Figure 9a shows the average activation of the model layer 6 ventrolateral prefrontal cortical cells at the time that errors were assessed, as well as the output threshold (see Section 5.1 for discussion of the threshold) .

Figure 9b plots the observed likelihood that a previously presented item (i.e., an item other than the one presented on the screen during the current inter-onset interval, or IOI) was rehearsed during each rehearsal set (i.e., IOI). To obtain this measure, a re-analysis of the Tan & Ward (2000) data was conducted. For each rehearsal set, the percentage of trials in which each item was recited was summed. Within the reported data, these recitations were pooled across overt rehearsals and the reading aloud of the concurrently presented item. In order to exclusively identify self-initiated rehearsals, the recitations of the currently presented item were subtracted from that sum. The result demonstrates that rehearsal performance was load-dependent; namely, subjects were far less likely to rehearse items during rehearsal periods near the end of a list, when more items had been previously presented.

Figure 9c plots the probability that a given item was recalled relative to its presented order. This is the standard serial position performance curve, termed *nominal serial position* in Tan & Ward (2000).

To derive Figure 9d, the last rehearsal set in which each item was read aloud or rehearsed was calculated and performance was averaged across the items in each rehearsal set. This measure provided a rough estimate of performance versus time of last vocalization.

Figure 9e plots the last time each item was vocalized, either during a rehearsal or presented on the screen. The items, and their associated probability of recall, were arranged on the x-axis based upon this order of last presentation (termed *functional* order in Tan & Ward, 2000). While the information contained in Figures 9d and 9e are similar, Figure 9e describes probability of recall based upon the *order* in which items were last rehearsed, whereas Figure 9d describes the probability of recall based upon a coarse measure of how long ago (i.e., *time*) they were last rehearsed.

Figure 9f shows the difference between the set in which an item was last rehearsed and the set in which it was first presented. As can be seen, LIST PARSE simulations closely fit all of these experimental data as emergent properties of working memory network dynamics (described in Sections 4.1.2 and 5.10 below).

Figure 9g illustrates how the model responds to paradigms that allow distracters, notably the delayed free recall and continuous-distracter free recall paradigms. To facilitate the comparison with immediate free recall, the solid curve in Figure 9g again shows the model simulation of immediate free recall from Figure 9c. It has long been known that, under appropriate experimental conditions, a distracter-filled post-list delay causes a significant reduction in recency effects (Glanzer, 1972; Glanzer & Cunitz, 1966; Postman & Phillips, 1965). The dotted curve in Figure 9g displays the serial position curve that the model generated during delayed free recall, when a delay period of 12.5 seconds (half the list presentation time in this protocol) was added between list presentation and recall during which no rehearsals were allowed to take place. Note the significant decrease in the recency gradient. In the continuous-distracter paradigm, distracters are included between the presentations of items in the list, in addition to during the post-list delay (Bjork & Whitten, 1974; Davelaar et al., 2005; Tan & Ward, 2000; Tzeng, 1973). Given this variation, the recency portion of the curve largely returns, and often the primacy effect in the list is attenuated. The dashed-dotted curve displays the result of a simulation during which rehearsal is prevented both during a delay period of 12.5 seconds added after the list as well as during the shorter ISIs between item presentations. See Sections 3.12 and 5.10 for further discussion of both of these effects and the corresponding experimental paradigms.

3.10. Differences Between Immediate Free and Serial Recall Performance. Although serial position performance curves for the ISR and IFR paradigms tend to have the same general bowed shape (Figures 6, 7 (circles), and 9c), there are obvious differences. The IFR serial position curve tends to have a more rounded shape and a substantially longer recency gradient than does ISR curve, regardless of presentation modality (Figure 9c). Overall list performance is better for IFR with the improvement most evident for the end of list. This occurs because the ISR task additionally penalizes order errors, but it occurs even when the same presentation schedules are used and only item errors in both ISR and IFR tasks are considered (Klein et al., 2005).

These serial position curve differences have been taken as evidence that different mechanisms underlie the two tasks (e.g. Tan & Ward, 2000). However, as discussed previously, experimental confounds exist between ISR and IFR tasks that may cause such differences even if they use the same working memory circuitry. Perhaps most importantly, different rehearsal

strategies are often used in the two protocols (Rundus, 1971; Tan & Ward, 2000). Serial recall rehearsal almost always proceeds from the beginning of the list and proceeds in order of presentation during each IOI. Free recall rehearsal can be highly variable. The LIST PARSE model suggests how these differences can arise through small variations in selection mechanisms between the two tasks, which otherwise share similar working memory representations.

For long IOIs, subjects frequently follow a serial recall rehearsal strategy, starting from the list beginning and going as far as possible in order. For short IOIs, rehearsal frequently begins with the most recently presented items. In addition, it may be possible to read-out the last few items from a separate "transient sensory memory" at the beginning of the IFR recall period, thereby increasing the recall probabilities of these final items which tend to be poorly recalled in an ISR task where such a strategy is impossible. The LIST PARSE model simulates properties of immediate free recall, delayed free recall, and continuous-distracter free recall without using such a transient sensory memory.

Klein, Addis, & Kahana (2005) provide additional evidence for such a viewpoint: Conditional response probabilities as a function of lag for a single trial are remarkably similar in both tasks. These conditional probabilities are defined as follows: Given recall of an item, j, compute the likelihood that the next item selected will be the item which was presented k items before or after j. This similarity holds, in particular, when subjects know that the ordering of items will remain the same over multiple trials, thereby encouraging serial rehearsal strategies in the IFR task. At present, LIST PARSE does not account for the recall order in the IFR task. Rather, the model explains the serial position curves for the recall period; namely, which items are recalled, but not their relative ordering.

3.1l. Evidence for Multiple Memory Stores. Davelaar, Haarmann, Goshen-Gottstein, Ashkenazi & Usher (2005) have proposed that dual-memory stores exist in which representations of episodic context in medial temporal lobe, or "episodic long-term memory", are associated with temporarily active learned representations of lexical/semantic categories in lateral prefrontal cortex, or "short-term store". This is a recent example of a dual-store account of free recall performance first introduced by Atkinson & Shiffrin (1968), in contrast to recently more popular single-store accounts of working memory (e.g., Ward, 2002; Howard & Kahana, 2002; see Section 6.1). A dual-store account was supported by Davelaar and colleagues largely on the basis of four dissociations in performance on two paradigms, the IFR and the continuous-distracter paradigm (an IFR protocol with distracter tasks occupying the ISIs and a post-list period before recall begins): output-order effects on serial-position functions (Dalezman, 1976; Whitten, 1978), dissociations due to amnesia (Carlesimo, Marfia, Loasses & Caltagirone, 1996), negative recency effects (Craik, 1970), and output-position effects on lag recency (Howard & Kahana, 1999; Kahana, 1996).

The LIST PARSE model supports the idea that at least two stores likely exist, but not necessarily in the manner that Davelaar et al (2005) characterized. A "short term store" is not explicitly modeled in the current LIST PARSE model presentation, but the concept is roughly equivalent to the concept of a "transient sensory memory," as distinct from a cognitive or motor working memory. This modality-specific memory source is postulated to explain modality effects (e.g., the pre-categorical acoustic store) and some of the differences between serial recall and IFR serial position curves. LIST PARSE demonstrates, however, that a list that is stored by the cognitive working memory can itself have a recency gradient (Figures 9 and 12c below). Indeed, the cognitive working memory is the sole basis for the shape of all the simulated serial position curves with the exception of the auditorily presented list in Figure 6a. Moreover, many

of the recency gradients that are found in free recall data are longer than the postulated span of a transient sensory memory. Lower-level sensory-specific areas in the parietal and temporal lobes may support such limited memory stores.

LIST PARSE working memory stores, both cognitive and motor (Figures 1 and 2), are also activity-based memories. However, entry into these memories is selective, and item activation, either individually or in sequences, can persist for extended periods of time. They can be affected by distracter tasks that prevent rehearsal, or by active displacement by subsequent items. These working memory stores are the foundation of almost all simulations in this paper, behave similarly to temporal distinctiveness models of free recall (discussed below in Section 7.1), and are proposed to involve lateral prefrontal cortex. While it may seem counter-intuitive to think of immediate recall of up to 20-item lists as a working memory phenomenon, performance in these tasks involves rehearsing throughout the task to keep as many items as possible simultaneously "in mind", a trademark of working memory performance. True "long-term memory" representations, which enable items (or sequences of items) to be restored to active status, are contained in learned associative weights from list chunks in layer 2/3 to item activation patterns in layers 4-6 (in connections $C_j \rightarrow Y_i$ in Figure 2 and Appendix equation (3)). These learned weights encode strengths that are similar to the working memory activations from which they were learned (Section 2.6). The data simulations in this article do not incorporate these STM-LTM interactions, but rather focus on how many data patterns can be replicated using properties of working memory that are not yet embodied in alternative models.

3.12. Influence of Distracters. LIST PARSE qualitatively explains a wide range of data about how distracters influence learning and recall. During free recall, placing an extended distracter task after list presentation, but before recall, preferentially and strongly attenuates recency effects. This effect is illustrated by comparing the solid and dotted curves in Figure 9g, which show the simulated serial position curves for immediate and delayed (with rehearsal prevented) free recall. However, if distracter tasks are additionally placed in the ISIs during list presentation (called the "continuous-distracter paradigm"), then the recency portions of the curves partially return. This latter property is called the "long-term recency effect" (Bjork & Whitten, 1974; Davelaar et al., 2005; Tan & Ward, 2000; Tzeng, 1973). It is illustrated by comparing the dotted and dashed-dotted curves in Figure 9g, which show the simulated serial position curves for delayed and continuous-distracter free recall, respectively. The mechanisms leading to the significant reduction in recency for the delayed free recall simulation and the partial reinstatement of recency in the continuous-distracter free recall are discussed conceptually below, and mechanistically in Section 5.10.

Distracter tasks are likely to have several effects. First, they eliminate "transient sensory memory" if the distracter task requires the use of this modality-specific memory store or introduces a sufficient delay before recall. Second, they prevent rehearsal in working memory while the distracter is used. As discussed in Sections 4.1, 5.10, and 7.1 and shown in Figure 12c below, in the absence of rehearsal, the presentation of a long list naturally leads to a bowed gradient with extended recency. The ability to rehearse during a list tends to preferentially promote primacy, due to increased rehearsals of early list items, and to diminish recency due to competition with these more active early list items. These effects are magnified when the stronger LTM representations of repeatedly rehearsed items boost these early items, and thus competitively reduce recency item representations within the self-normalizing network; cf., the comparison between the transient memory span and the immediate memory span in Section 2.6. In addition, when free recall of a long list is allowed to progress unfettered by distracters, the

transient sensory memory can allow the subject to begin recall by reading out the last few items in the list, resulting in enhanced recency. The inclusion of a filled retention interval after list presentation can eliminate the transient sensory memory, thereby reducing recency. Diminution of recency would also be expected by requiring recall to begin with the start of the list, because this also eliminates the influence of the transient sensory memory (see Section 5.7; Dalezman, 1976).

The continuous distracter protocol includes distracters during the ISIs, and thereby eliminates the tendency toward enhanced primacy due to inter-trial rehearsals in working memory. The lack of rehearsals keeps the LTM representations that build up from such repeated rehearsals from selectively promoting primacy items in a similar manner. Also, the rapid shifting back and forth of distracter tasks and item presentations in the continuous-distracter task raises the possibility that the volitional signals that must be strongly engaged to allow for the good retention of items in working memory are weakened. The weak retention (or coding) of primacy items during list presentation serves to enhance the retention of later items, and thus recency (see Figure 12c below), due to self-normalizing dynamics. Because these working memory representations drive learning and list-chunk-mediated read-out of long-term memories, these recency effects can later be reinstated from long-term memory. These experimental paradigms are reviewed in greater detail by Davelaar et al. (2005) and Tan & Ward (2000).

3.13. Similarities between Spatial and Non-spatial Processing. Another important question concerns whether spatial and non-spatial sequential information is processed and retained similarly in working memory, given evidence for distinct processing localizations even within granular prefrontal cortex (see Section 2 and Figure 1). Available evidence suggests the use of shared neural mechanisms, even in distinct cortical cell populations. For example, Jones, Farrand, Stuart, & Morris (1995) conducted a spatial serial recall task in which visual target locations were sequentially presented and subjects subsequently recalled the locations in their presented order. The authors reported strikingly similar results to those seen in the non-spatial (e.g., auditory or visual object serial recall) literature, including similar serial position bows, list length effects including similar measures of span capacity, interference effects from items and actions in different modalities, and retention interval effects. Agam, Bullock and Sekuler (2005) sequentially presented spatial targets and instructed subjects to manually imitate the sequence. The authors reported that subjects not only made errors that closely resembled the bowed (1-item recency) shape of the serial position performance curves, but the majority of sequential errors were transpositions of neighboring movements, mimicking a central property of the error-type distribution of non-spatial ISR (e.g. Henson et al., 1996). These results from humans in multiple spatial and non-spatial tasks are consistent with the Grossberg (1978a, 1978b) prediction that all working memories, in all modalities, obey similar rules, in order to be able to stably learn and remember list categories, and use them to recall from long-term memory the sequence information that is stored in short-term working memory.

3.14. Neurophysiological Data about Working Memory Storage and Performance. There has been an explosion of neurophysiological data concerning single-cell neural properties related to the storage of event sequences in working memory and their serial performance through time. Experimental data (Figure 10a) that strongly support predictions of Item and Order working memory models, (Grossberg, 1978a, 1978b) and that are simulated by LIST PARSE (Figure 10b), were reported by Averbeck et al. (2002, 2003a, 2003b). These authors did extra-cellular recording from the areas near, but outside the depths of, the caudal portions of the principal sulcus of macaque monkeys during the performance of a sequential motor task. In this task,

monkeys used an X-Y joystick to copy concurrently presented geometrical shapes (triangle, square, inverted triangle, trapezoid) on an LCD screen in a prescribed order. Copying proceeded counter-clockwise starting at the top middle of each shape. The recorded cell responses were pooled on the basis of the movement in the sequence with which their firing pattern most correlated. The population response for each movement in the series is shown in Figure 10a.

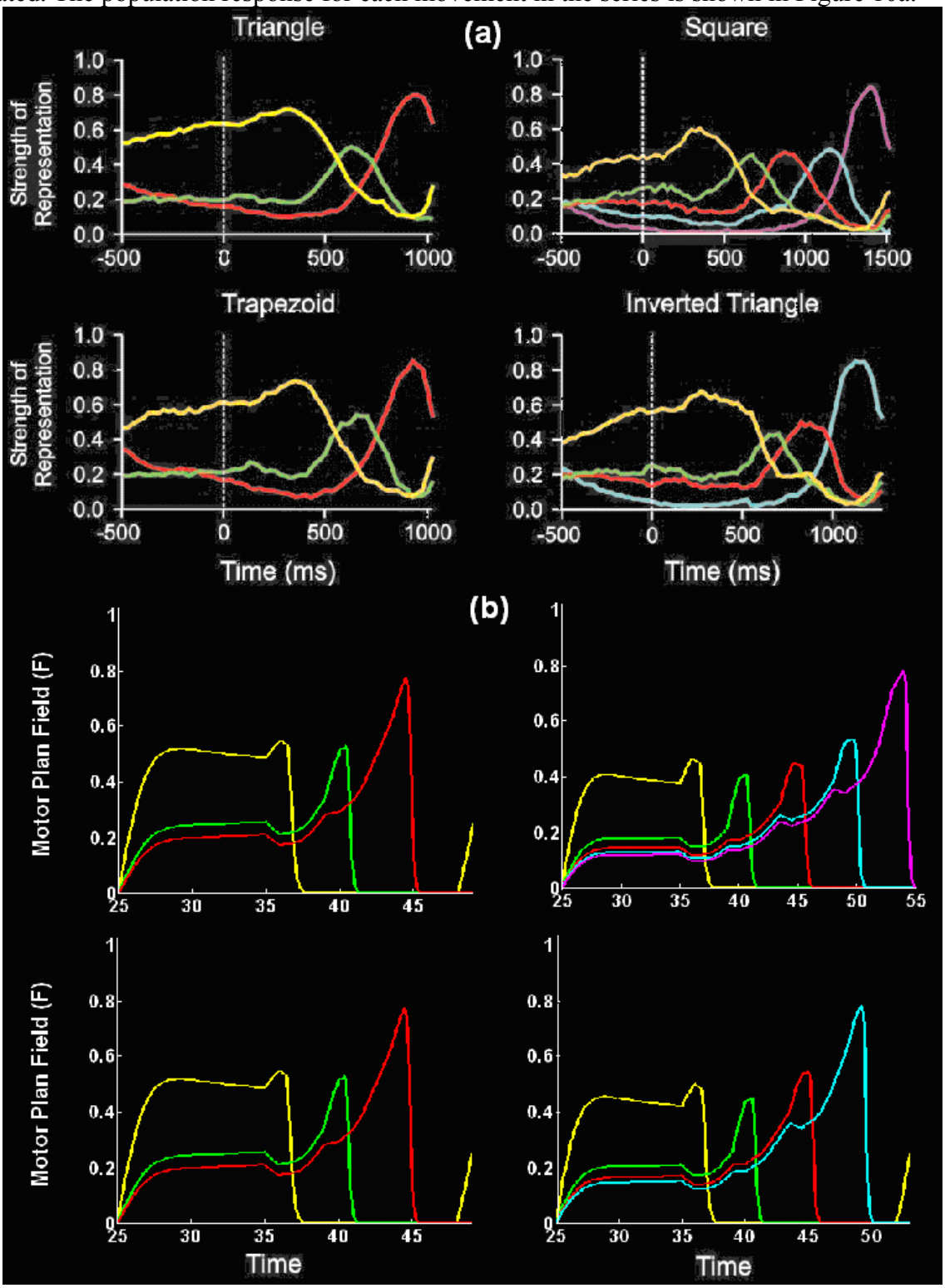


Figure 10. (a) Plots of relative strength of representation (a complex measure of cell population activity, as defined by Averbeck et al., 2002) vs. time for four different produced geometric shapes. Each plot shows the relative strength of representation of each segment for each time bin (at 25ms) of the task. Time 0 indicates the onset of the template. Lengths of segments were normalized to permit averaging across trials. Plots show parallel representation of segments before initiation of copying. Further, rank order of strength of representation before copying corresponds to the serial position of the segment in the series. The rank order evolves during the drawing to maintain the serial position code. (b) Simulations of item activity across the motor plan field of the LIST PARSE model for 3, 4, and 5 item sequences vs. simulation time. In both (a) and (b), curve colors correspond to representations of segments as follows: yellow, segment 1; green, segment 2; red, segment 3; cyan, segment 4; magenta, segment 5. (Reproduced with permission from Averbeck et al., 2002).

Four phases of the Averbeck et al. (2002; Figure 10a) curves should be noted, all of which support predicted Item and Order working memory properties: First, a primacy gradient is established, so that greater relative activity corresponds to earlier eventual execution, in the sequence during the period prior to the initiation of the movement sequence (period -500 to 400ms). Second, contrast enhancement of the activity pattern favors the item to be performed by causing a greater proportional activation of the first item representation prior to first item performance (period ~100 to 400ms). Next, reduction of an item representation's activity occurs just prior to its performance (inhibition of return) along with relative enhancement of the representation of the next item to be performed such that it becomes the most active item prior to its execution (period ~400ms to near sequence completion). Finally, there is a possible re-establishment of the gradient just prior to task completion. Similar dynamics are exhibited by the activation levels of the motor plan storage cell populations, F_i, of the LIST PARSE model (Figure 10b; Appendix equation (9)).

Using a paradigm that dissociated sequential arm movements from the corresponding on-screen cursor movements, Mushiake et al. (2006) replicated each of these parallel planning properties in lateral prefrontal cortex with only one small exception, an over-representation of the final item during the preparatory period. The authors found that it was the resulting on-screen cursor movements, rather than the actual arm movements creating them (properties that were confounded in the Averbeck et al. (2002, 2003a, 2003b) studies), that were predicted by dorsolateral prefrontal activations. In contrast, recordings in primary motor cortex showed selectivity for arm movements as opposed to the cursor, and failed to show significant parallel planning during the sequence preparatory period. Instead, only serial rises in representation were recorded just before and during each planned arm movement.

The Averbeck et al. (2002, 2003a, 2003b) studies are part of an explosion of neurophysiological data concerning single-cell neural properties that are related to sequential order. A more comprehensive LIST PARSE architecture, which will be outlined in Section 6, provides a qualitative explanation of many of these cell properties. The concepts in this architecture build upon the mechanisms and properties of the more circumscribed LIST PARSE model that is quantitatively defined and used to explain and simulate its currently targeted range of cognitive and neural data. This architecture will be quantitatively described in subsequent articles.

The additional cell properties that the architecture can accommodate include such factors as repeated items in a sequence and the ordinal position of an event in a sequence, leading to a theory of how context-sensitive sensory and motor sequence chunks represent past events, and control the performance of future sequences of events in a manner that is sensitive to their ordinal position. Here are some of the cell properties that have been reported:

Selectivity for ordinal position. For example, a cell may respond to any target in the initial, middle or final ordinal position in a sequence, or the first, second, third, etc. positions. These cells are said to exhibit *rank order*, also called temporal selectivity or position specificity (Averbeck et al., 2003a; Barone & Jacobs, 1989; Funahashi et al., 1997; Inoue & Mikami, 2006: Kermadi & Joseph, 1995; Ninokura et al. 2004).

Selectivity to *conjunctive coding* of item and ordinal position. Such a cell may respond to a specific target presented in a specific ordinal position (Averbeck et al., 2003a; Funahashi et al., 1997; Inoue & Mikami, 2006; Ninokura et al., 2003, 2004; Shima & Tanji, 1998, 2000).

Selectivity for sequence context. Here a cell may respond to an item differently when it is part of one, or many, specific sequences rather than when it is isolated (Funahashi et al., 1997; Kermadi & Joseph, 1995; Mushiake et al., 2006; Ninokura et al., 2003).

Selectivity for individual sequential orderings. Such a cell may respond when a specific item combination or ordering, A-C-B or B-A, is presented. This property is sometimes called pair-dependent (Averbeck et al., 2003a; Funahashi et al., 1997; Kermadi et al., 1993; Kermadi & Joseph, 1995; Mushiake et al., 2006; Ninokura et al., 2003, 2004).

Selectivity to instruction or item repetition (Kermadi & Joseph, 1995; Shima et al., 2007).

Selectivity to abstract sequence category or type, such as when sequences require a repetition, or have a particular movement combination as a sub-component, regardless of the specific movements (Shima et al., 2007).

In addition, one finds in these experiments some cells whose selectivities include no clear sequential or positional dependence, but are selective for properties such as input stimuli (e.g. Inoue & Mikami, 2006; Kermadi & Joseph, 1995; Ninokura et al., 2004), responses (e.g. Kermadi & Joseph, 1995; Takeda & Funahashi, 2002), task goals (Mushiake et al., 2006), performance metrics (Averbeck et al. 2003a), rewards (e.g. Kermadi & Joseph, 1995; Mushiake et al., 2006), or errors (Kermadi & Joseph, 1995).

Almost all of these studies (Barone & Joseph, 1989; Funashi et al.,1997; Inoue & Mikami, 2006; Kermadi et al., 1993; Kermadi & Joseph, 1995; Mushiake et al., 2006; Ninokura et al., 2003, 2004; Shima et al., 2007) explored sequential effects using only two or three item sequences. Nearly all of these conceptually separable properties become indistinguishable when applied to lists of such short length, which further complicates concise explanation or interpretation. In particular, the first item in a two item list is simultaneously in absolute position one, a sequence beginning or initiation, and the "earlier" position relative to every other item or subsequence. An item is in the second position of a three item list could indicate specificity for the absolute second item, a "middle" item, an initial item for a two item sub-list, a final item for a two item sub-list, a "later" relative position when compared to the first item, and an "earlier" relative position when compared to every item other than the first.

Both the Averbeck et al. (2002, 2003a, 2003b) and Mushiake et al. (2006) studies manage to side-step many of the single-cell complexities, by assessing measures of the amount of total information contained within the aggregate neural population code that is predictive of each movement for each time bin throughout a trial. This is accomplished by retroactively finding the

correlation between times that a cell's response rises or falls and the observed posterior probability that a given response was made.

4. LIST PARSE model. This section provides an intuitive functional explanation of the LIST PARSE model processes and their interactions. The underlined portions of Figure 1 are modeled herein, and Figure 2 depicts a circuit diagram of modeled processes. Subsequent figures (Figures 11 and 14) provide more specific circuit diagrams of the motor aspects of the model. A more detailed description of how the simulations (Figures 6, 7, 8, 9, 10, and 16) and activity profiles of the model (Figures 12, 13, and 15) were produced and what dynamics of the model produce them is presented in the next section and in the Appendix. Mathematical equations and parameters for all modeled components are also provided in the Appendix.

The Baddeley (1986) model of working memory has remained an effective framework for organizing cognitive research since its proposal, and has received significant support from the rapidly expanding functional imaging experimental literature (e.g., Braver et al., 2001; Chein & Fiez, 2001; Marshuetz et al., 2000; Paulesu et al., 1993; Poeppel, 1996; Smith et al., 1998). The Baddeley model proposes that working memory function can best be described as a central executive controlling two subsystems that are specialized by the information they represent; namely, the *phonological loop* and *visuospatial sketchpad*.

The LIST PARSE model provides a computational account of the phonological loop consisting of a BA22/IPL → cognitive working memory (ventrolateral PFC) → motor working memory (dorsolateral PFC) → ventrolateral premotor cortex and anterior insula → BA22/IPL loop for covert rehearsal (see Figure 1). It also provides an account of the visuospatial sketchpad, consistent with the premotor theory of attention (e.g. Rizzolatti, Riggio, Dascola & Umilta, 1987), consisting of high-level dorsal and ventral visual streams → cognitive working memory (principal sulcus in the monkey) → motor working memory (dorsolateral PFC and SEF) → FEF → dorsal and ventral visual streams (see Figure 1). LIST PARSE also elaborates how these sub-systems are controlled by volitional executive functions such as non-specific gain control mechanisms and task performance gating.

4.1 Cognitive Working Memory

4.1.1. Unitized Category Inputs to Layers 4 and 6. The inputs I_i to the LIST PARSE model represent familiar object or event categories that have been unitized through previous learning at lower stages of the processing hierarchy; see Section 2.1 and Figure 1. These categories are called *items*, and are designated as Item Category Inputs (*I*) in Figure 11a.

The model assumes that occurrences of these unitized items cause stereotyped transient, or pulse, inputs to the LIST PARSE working memory. All the inputs I_i to the cognitive working memory network in Figures 2, 11, and 12a are delivered to both layer 4 and 6 (Appendix equations (2) and (3)). The inputs have the same fixed duration (1 simulation time unit), the same magnitude (.1), and are separated in time by the IOIs of the task that is being modeled with one simulation time unit equal to 100ms of experimental time. These assumptions allow the working memory to code temporal order without bias.

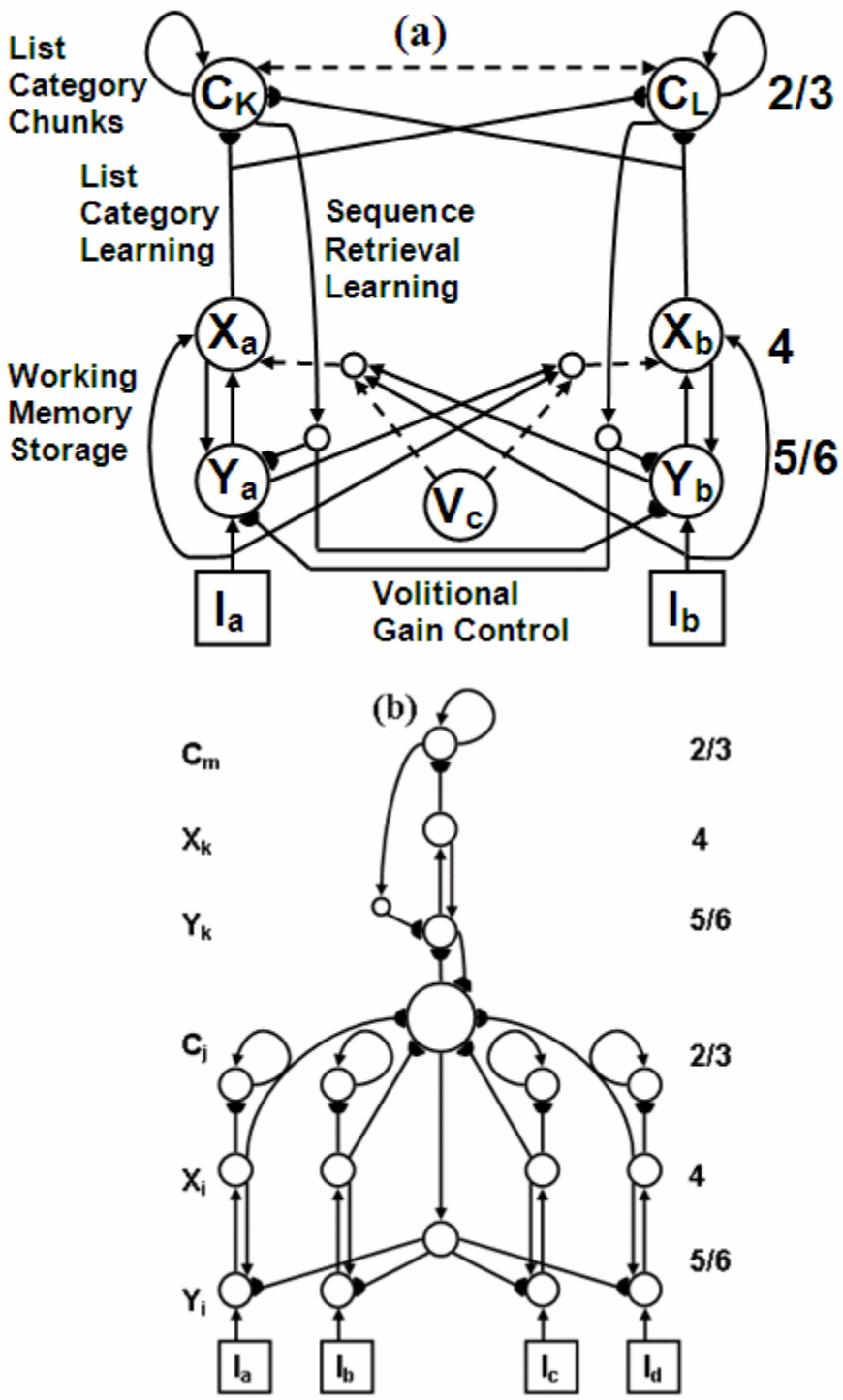


Figure 11. (a) Two columns (a and b) of the cognitive working memory circuit. Given a temporally separated sequence of inputs (*I*), layers 4 (*X*) and 6 (*Y*) form a positive feedback loop that sustains an activity gradient that codes both item and order information. This stored pattern can be modulated by a volitional gain control (*F*) which serves to alter the level of cross-columnar competition within this loop. Stored item subsequences (*X*) learn to activate layer 2/3 list category chunks (*C*) via a learned layer 4-to-2/3 adaptive filter. These list categories (*C*) bias the working memory pattern (in *Y*)

and learn to re-institute (retrieve) their coded pattern through learned top-down expectations (sequence retrieval learning) from layer 2/3-to-5/6. (b) Interactions between two processing levels in a hierarchical working memory. In the lower level, four simultaneously active items (X_i and Y_i loops) are unitized into a chunk (C) that codes for all four items through learned feedforward and feed-back connections. Only one chunk is shown, for simplicity. This larger chunk differs from the one-item chunks shown only in the number of inputs it receives and in the strength of its competitive connections with other chunks in layer 2/3. This unitized chunk serves as bottom-up input to a higher-level cognitive working memory as an item representation, where it too can be combined with other items into a unitized list chunk. Competitive interactions among the list chunks (C) and the inhibitory off-surround connections between layer 4 and 6 are omitted.

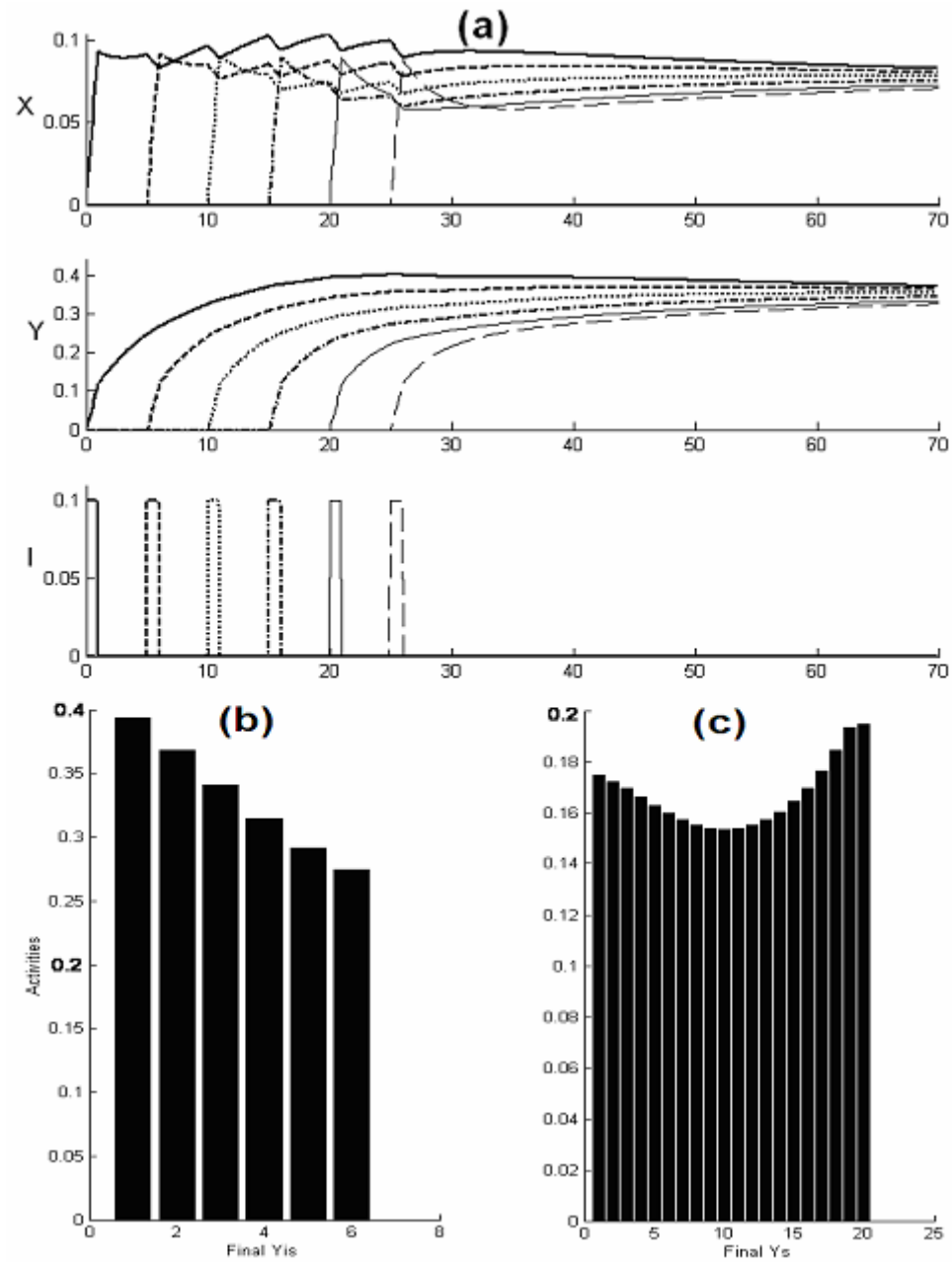


Figure 12. (a) Simulation of the development of a primacy gradient: activation across time of cells in layers 4 (X) and 5/6 (Y) of the cognitive working memory results from the presentation of a 6 input (I) list with 5 simulation time unit IOIs. Cells coding for each item channel are represented as follows: item 1, dark solid curve; item 2, dark dashed curve; item 3, dotted curve; item 4, dash-dot alternating curve; item 5, light solid curve; item 6, light dashed curve. This convention for displaying multiple stored item activations in parallel is maintained through the remainder of the paper. (b) Simulation of the same primacy gradient as in (a) across the layer 5/6 cells (Y) of the cognitive working memory 5 simulation time units after last item IOI when a sub-span list of 6 items. (c) Simulation of a bowed gradient (with extended recency) in response to a super-span list of 20 items presented with 5 simulation time unit IOIs.

4.1.2. Working Memory in Layers 6 and 4: Self-normalizing Competition Causes Primacy and Bowed Gradients. Temporal order information of item sequences is stored in working memory by recurrent on-center off-surround networks whose cells obey membrane equations (that is, obey shunting dynamics). Such a network occurs between model layers 6 and 4 (6/4) of ventrolateral prefrontal cortex and the depths of the principal sulcus; Variables X_i in layer 4, and Y_i in layer 6 represent the item corresponding to input I_i. See Figures 2, 11a, and 12a, and Appendix equations (2) and (3). Both item and order information of a temporally ordered sequence of n inputs I_1, I_2, …., I_n are maintained by the relative activities $Y_1, Y_2, \ldots, Y_n$, with correct temporal order stored by a primacy gradient wherein $Y_1 > Y_2 > \ldots > Y_n > 0$. This circuit, with 3 free parameters fitted to one task (Figure 7), plus noise and threshold values, simulates all serial recall and free recall data in Figures 6-9, with additional assumptions about selection strategy during rehearsal.

The shape of the activation gradient in cognitive working memory for short and long lists is determined as follows (see Figure 12). The shunting on-center off-surround dynamics of layer 4 partially normalize activity across the network in response to inputs I_i. The lack of competitive dynamics in layer 6 (Appendix equation (3)) causes these cells to behave as leaky integrators. Layer 6 cell activities, particularly for short lists, are therefore largely dictated by the length of time they integrate the activities of layer 4 cells. Early items integrate longer, so primacy dominates. Thus, a primacy gradient can be read out of the cognitive working memory layer 6 after the positive feedback loop between layers 4 and 6 has a chance to act.

As more items enter the working memory network in response to a long list, the activities of layer 4 cells decrease due to competitive self-normalizing dynamics, because approximately the same total network activation is spread across the larger number of active cells. Given that these layer 4 cells drive layer 6 cells, the leakage term in the layer 6 equation (Appendix equation (3)) causes the average activity level of layer 6 cells to drop too. When a sufficient number of items are stored, corresponding approximately to span list length in the ISR paradigm, the bottom-up pulse inputs to both layer 4 and layer 6 drive the incoming items to an activity level greater than at least one stored item, and a recency gradient begins. When the bottom-up input ceases, similar normalizing dynamics diminish this new item's representation so that, when the next item is presented it has greater activity than the last item and an extended recency gradient forms. Thus, the same dynamics that produce a primacy gradient for short lists can produce an extended recency gradient for long lists.

4.1.3. Volitional Control of Working Memory Storage by Basal Ganglia. The strength of working memory competition is controlled by an inhibitory gain parameter F (see Appendix

equation (2)). A volitional gain control signal, V_c, modulates the size of F and thereby controls the strength of the network's competitive interactions (Figure 2). A large F makes it harder to store items in working memory. Since V_c inhibits F, turning on cognitive volition facilitates storage of items in working memory, and therefore leads to a larger memory span; see Figures 2 and 11a. In particular, a small volitional signal V_c, and thus a large competitive gain F, makes it difficult for items to become significantly activated in layer 4, and layer 6 in turn, and for bottom-up inputs to enable recent items to have greater activation than those already present. A medium V_c, and thus a medium F, allows a few items to become significantly active, but limits the average activity level in layer 4, which leads to an early onset of recency, and thus a reduced ISR span. Finally, strong volition V_c, and weak competition F, allows many items to become active, and enables the accurate retention of their order information in a primacy gradient. One possible source of volitional gain control is the basal ganglia. See Brown, Bullock, and Grossberg (2004) for a detailed model of how the basal ganglia can gate thalamocortical dynamics.

4.1.4. Learned Bottom-Up Grouping of Item Sequences by List Chunks in Layers 2/3. Item sequences that are stored in working memory may be grouped through bottom-up learning from layer 4-to-2/3 into list chunks, or unitized sequence codes, in layers 2/3 (Figure 2, pathways labeled List Category Learning in Figure 11a, Appendix equation (6)). The list chunk network is defined by a recurrent on-center off-surround shunting network (Figure 2 and List Category Chunks in Figure 11a, Appendix equation (5)). Cells in the list chunk network come in multiple sizes, or scales, such that larger cells selectively code longer lists. In particular, larger cells in such a *masking field* can competitively "mask," or inhibit, the activities of smaller cells, more than conversely (Grossberg, 1978a; see also: Cohen & Grossberg, 1986, 1987; Grossberg & Myers, 2000; Nigrin, 1993; Page, 1994).

LIST PARSE simulates how a laminar masking field circuit in layer 2/3 can learn list chunks, for lists of variable length, in response to dynamically evolving patterns of activity in layers 4/6. In particular, stored activities X_i in layer 4 project to the list chunks C_k in layer 2/3 via an adaptive filter, which comprises the pathways labeled List Category Learning in Figure 11a. Excitatory and inhibitory feedback within layer 2/3 in response to these input patterns from layer 4 contrast-enhances the activity of some layer 2/3 cells while inhibiting the activities of all other layer 2/3 cells. The winning layer 2/3 cells drive learning by the adaptive weights within the layer 4-to-2/3 pathways. After learning occurs, different stored sequences in layers 4/6 activate different masking field list chunks.

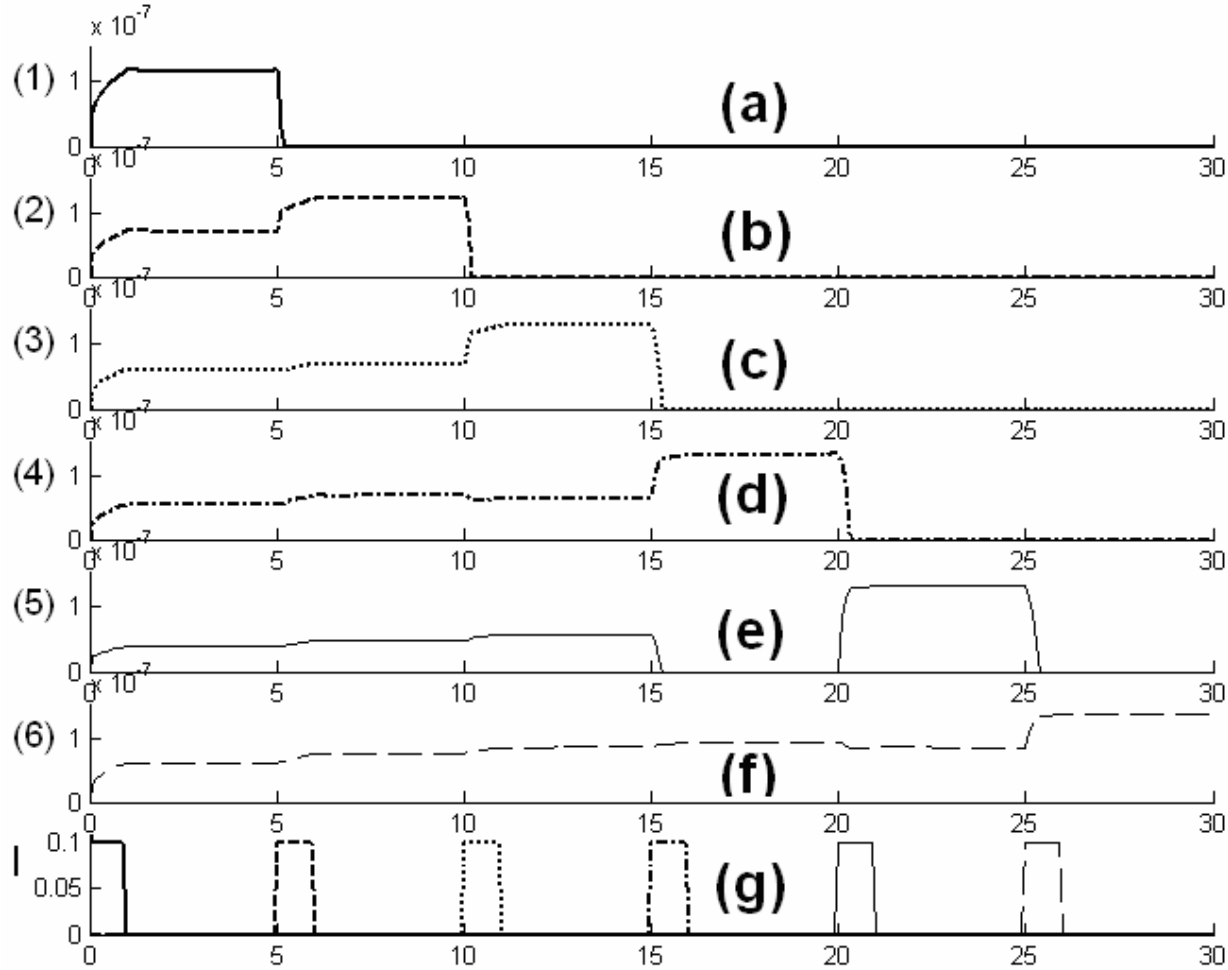


Figure 13. List chunk activations across time of layer 2/3 cells (*C*) coding for different length subsequences of a presented list of 6 items, with 5 simulation time unit IOIs. (a) Connected only to the first item in the sequence. (b) Connected only to the first 2 items. (c) Connected only to the first 3 items. (d) Connected only to the first 4 items. (e) Connected only to the first 5 items. (f) Connected to all of the 6 items in the presented list. (g) The inputs to the cognitive working memory.

The masking field dynamics in layers 2/3 are illustrated in Figure 13 and include the following properties:

Priming: Partial activation of a list chunk occurs when the items in working memory form a subset of the list that this chunk has learned to code.

Sequence selectivity: Strong activation of a list chunk occurs when its preferred item sequence is active in working memory. This most predictive chunk can competitively suppress list chunks corresponding to shorter lists.

Long list potency: Sequence selectivity implies that an active list chunk can be competitively suppressed when a chunk corresponding to an even longer list is activated. As a result of this property, a "resonant transfer" can occur from shorter to longer list chunks as a sequence of items is stored through time in working memory (Grossberg & Myers, 2000).

Taken together, these properties show that masking field list chunk activities reflect the "amount of evidence" for the list that they code, notably the best prediction that can be based upon the current temporal context.

Priming is a special case of the sequence selectivity of a masking field. As remarked in Section 2.4, it prevents a list chunk that corresponds to a longer list (for instance, a word like MYSELF) from suppressing activation of a cell that has learned to code a subsequence of this list (for instance, the words MY, SELF, or ELF), when only that subsequence of items is presented. In contrast, long list potency comes into play when such a longer sequence chunk is fully activated by its bottom-up input sequence (e.g., MYSELF). Then, and only then, can it inhibit subsequence chunks (e.g., MY), and thereby enable the chunk for MYSELF to learn its own meaning without interference from the subsequence chunks. With this property ensured, other crucial properties follow, such as the selective association of chosen list chunks with sensory representations that reflect the meaning of the sequence, with motor representations that control actions compatible with these meanings, and with rewards or punishments that may occur after these actions are carried out.

4.1.5. Top-Down Learning from List Chunks to Working Memory Items. List chunks gradually learn to reactivate the pattern of items in the cognitive working memory that activated them, so that their activation can control unitized read-out of a sequence of items into the cognitive working memory system. Top-down learning occurs from the emerging list chunks to the item sequences stored in working memory, via learned expectation pathways from layers 2/3-to-4/6 (Figure 2, pathways labeled Sequence Retrieval Learning in Figure 11a, Appendix equation (8)) Top-down learning helps to stabilize the bottom-up learning of list chunks via general ART properties (Carpenter and Grossberg, 1991; Grossberg, 1980, 2003a). Top-down learning also enables selective activation of a list chunk (e.g., of a learned word) to read-out its learned weights to reactivate items in working memory through long-term memory retrieval, and thereby enable their fluent execution via the Motor Working Memory (Figure 2, Appendix equation (9).)

Top-down read-out from list chunks can also bias an already active working memory activity pattern in layers 4/6 towards storing a more extended primacy gradient. The tendency of the serial position curve toward primacy with increasing post-presentation delays, even in cases where rehearsal is deemed impossible (e.g., Knoedler et al., 1997), may be explained by a learned primacy gradient in the LTM feedback from activated list chunks to the working memory. As noted in Section 2.6, this top-down LTM read-out property helps to distinguish the transient memory span, or TMS, which operates without significant LTM read-out, from the immediate memory span, or IMS, which operates with LTM read-out.

In summary, the LIST PARSE model proposes how the temporal order of item sequences may be stored in working memory via recurrent on-center off-surround interactions in the deeper layers 6 through 4, how learned list chunks may be learned in the superficial layers 2/3 using a bottom-up adaptive filter from layers 4/6-to-2/3, and how learned top-down expectations from layers 2/3-to-5/6 enable a list chunk can to retrieve into short-term working memory the temporal order information across items that is stored in its long-term memory traces.

4.1.6. Higher-Order Cognitive Invariants: Chunks of Chunks. Figure 11b shows how the list chunks in layer 2/3 of one region of prefrontal cortex can be the inputs to layers 6 and 4 a higher cortical area, whose "items" are list chunks and whose "list chunks" are chunks of chunks. In particular, if such a higher-level list chunk network is a masking field, then it can encode a much larger temporal context, and will be reset much more slowly than individual item

representations as recall proceeds. Such a hierarchical organization allows higher-level chunks to individually represent increasingly large temporal contexts, such as sequences of words, and thereby to more reliably, and more deeply into the future, predict meaning and action.

4.1.7. Homologous Laminar Cortical Circuits for Spatial and Temporal Processing? LIST PARSE predictions about different roles for the lower and upper layers of ventrolateral prefrontal cortex and the depths of the principal sulcus (Figures 1, 2, and 11b), go beyond known neurophysiological data. If these predictions are confirmed, then a *temporal* working memory and list chunking network may be viewed as a variation on the more *spatial* circuits within the visual cortex.

Indeed, both cortices have highly granular six-layered cytoarchitectures, with a well-developed granule cell layer IV (Barbas & Pandya, 1989; Barbas & Rempel-Clower, 1997; Calloway, 1998). Moreover, in visual cortical area V1, recurrent layer 4/6 on-center off-surround interactions seem to play an important role in controlling the development and maintenance of cortical maps of ocular dominance and orientation tuning (e.g., Grossberg & Seitz, 2003; Grossberg & Williamson, 2001; Kayser & Miller, 2002; Olson & Grossberg, 1998). Layer 2/3 of V1 includes depth-selective complex cells (e.g., Hubel & Wiesel, 1968; Poggio, 1972; Poggio & Fischer, 1977; Poggio & Talbot, 1981; Smith et al., 1997), which have been proposed to derive their receptive field properties via an adaptive filter from layers 6/4-to-2/3, including grouping effects due to horizontal interactions within layer 2/3. Feedback interactions from layers 2/3-to-6/4 are predicted to help to select the monocular cells that are consistent with winning binocular cells in layers 2/3. Inputs from layers 2/3 of cortical area V1 to layers 6 and 4 of V2 lead to the formation of long-range perceptual groupings in layers 2/3 of V2, which may be interpreted as higher-order "list chunks" of V1. Several articles model these filtering and grouping properties and summarize relevant psychophysical, anatomical, and neurophysiological data (Cao & Grossberg, 2005; Grossberg & Raizada, 2000; Grossberg & Swamanathan, 2004; Grossberg & Yazdanbakhsh, 2005; Raizada & Grossberg, 2003).

Such results build upon a general anatomical framework of cortical organization that was proposed by Felleman and Van Essen (1991): Feedforward projections tend to originate primarily in the supragranular layers of a lower cortical area and to terminate in layer IV of a higher area, while feedback connections tend to originate primarily in the infragranular layers of a higher area and to terminate in layer I, and possibly layer VI, of the lower area. Barbas and colleagues have extended and refined this laminar framework (e.g., Barbas & Rempel-Clower, 1997; Dombrowski, Hilgetag & Barbas, 2001; Hilgetag, Dombrowski & Barbas, 2002; Rempel-Clower & Barbas, 2000) and similar hierarchical connection patterns have been found between distinct areas of frontal cortex and inferotemporal cortex (Barbas & Rempel-Clower, 2000) and posterior parietal cortex (Goldman-Rakic, 1987).

4.1.8. Homologous Working Memory Storage and Visual Imagery and Fantasy? A finer prediction of the homology between visual cortex and prefrontal cortex concerns volitional control. When cognitive volition V_c turns on, it inhibits the inhibitory interneurons in the cognitive working memory, thereby enabling item storage (Figures 2 and 11a). When volition is shut off, inhibitory gain increases. Inhibition then overwhelms excitation and the stored activities collapse, or are reset. The predicted anatomical site of this volitional inhibitory signal is at the inhibitory interneurons within layer 4. This prediction is consistent with an earlier prediction that a similar volitional circuit occurs in the visual cortex.

In visual cortex, top-down expectations usually provide attentional modulation of bottom-up inputs. They do so via a top-down, *modulatory* on-center, off-surround network that has its effect

on layer 4 cells (Grossberg, 1999a, 2003a). The modulatory on-center can sensitize target cells to respond more vigorously and synchronously to attended visual feature combinations. Increasing volitional gain inhibits the inhibitory interneurons, and thereby converts the modulatory on-center into one that can drive suprathreshold activation of its target cells via a top-down expectation. This volitional mechanism has been predicted to enable top-down expectations to generate suprathreshold conscious percepts of visual imagery and fantasy, rather than merely modulatory attentional feedback (Grossberg, 2000a). When this type of phasic volitional control over visual imagery and fantasy is replaced by tonic hyperactivity of the gain control source, hallucinations can occur that have many of the properties of schizophrenic hallucinations.

If these predictions about volitional control are supported, then they will provide another example whereby homologous mechanisms within a broadly used neocortical circuit design can carry out different functions: working memory storage in prefrontal cortex vs. visual imagery and fantasy in visual cortex.

This is not the only way in which the balance between excitatory and inhibitory signals in working memory and visual imagery circuits might be stored and reset. More experiments are needed to study the locus and action of the predicted nonspecific gain control mechanism, whose source is anticipated to be in the basal ganglia.

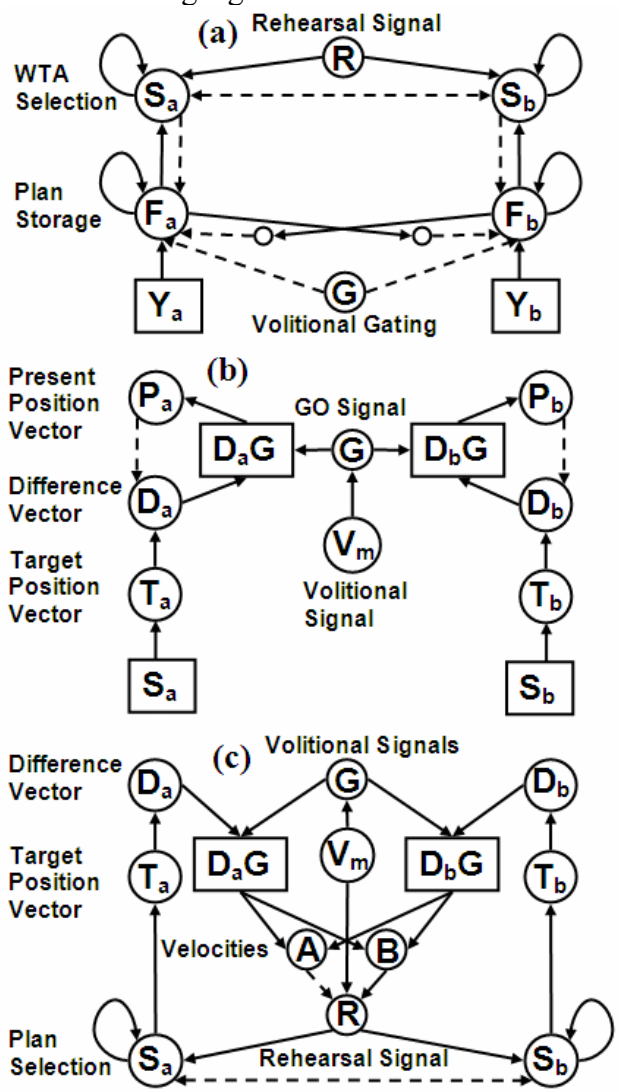


Figure 14. (a) Two columns in the motor working memory. Inputs from cognitive working memory (Y) are stored in the self-normalizing motor plan storage network (F). When a rehearsal signal (R) is active (above a zero threshold level), these parallel motor plans are instated in a plan selection layer (S). When a plan becomes super-threshold, it inhibits other plans, thereby preventing simultaneous execution of multiple plans, and sends self-inhibitory feedback to its representation in the motor plan storage network, thereby preventing perseveration. (b) Two columns in the trajectory generator circuit. Inputs from the plan selection field (S) of the motor working memory system establish a desired (target, T) configuration of the relevant motor effectors. The system maintains a representation of the present configuration (P) of these motor effectors. Trajectories are generated at a volitionally scaled rate (G) through a negative feedback loop that reduced the difference (D) between these two configurations. (c) The signals influencing the rehearsal signal (R). Selection of the next item in the plan selection field (S) is gated by the rehearsal signal. The rehearsal signal is activated by volitional signals (V) indicating that the circuit is in the recall phase of the task. The rehearsal signal is also modulated by a deceleration estimate ($B - A$) from slow (B) and fast (A) cells that time-average outflow velocity signals at different rates.

4.2 Motor Working Memory

4.2.1. Volitional Gating Between Cognitive and Motor Working Memory. The cognitive working memory activates cells in the motor working memory (Figures 2 and 14a, Appendix equation (9)). The items in motor working memory are well-learned motor plans that are associated through learning with the stimulus-derived items in the cognitive working memory. This learning process is not modeled here. Rather, it is assumed for simplicity that each cognitive item in layer 6 (with activity Y_i) can activate a corresponding motor item (with activity F_i). As a result, the spatial pattern of items stored in the cognitive working memory is transferred and stored in motor working memory. Read-out from motor working memory simulates the neurophysiological data in Figure 10, with performance controlled by the motor trajectory generators (Figure 14b, Appendix equations (15)-(18)) and the variable-rate volitional controller (Figure 14c, Appendix equations (19)-(21)).

Rehearsal of an item sequence from motor working memory is possible if each item is inhibited right after it is chosen for performance. Such self-inhibition, or inhibition of return, helps to prevent perseverative performance of a single item, as noted in Section 2 and further explained below. Reset of motor working memory does not, however, require reset of cognitive working memory. Thus, items stored in the cognitive working memory can be rehearsed multiple times, and recall of item lists can be restarted in mid-performance, even without LTM recall.

The same volitional GO signal, G, that gates *on* motor performance via the trajectory generator (Figure 14b, Appendix Equation (17)) also gates *off* read-out from the cognitive working memory to the motor working memory (Figure 14a, Appendix equation (9), term $\left(1 - f_4\left(G\right)\right)$). Volitional gating also helps to prevent perseveration: Motor performance can cause sensory feedback to the cognitive working memory via the environment without reinstating the performed item in motor working memory. Rehearsal can hereby influence performance, as it does in our simulations of cognitive data, without forcing pathological perseveration.

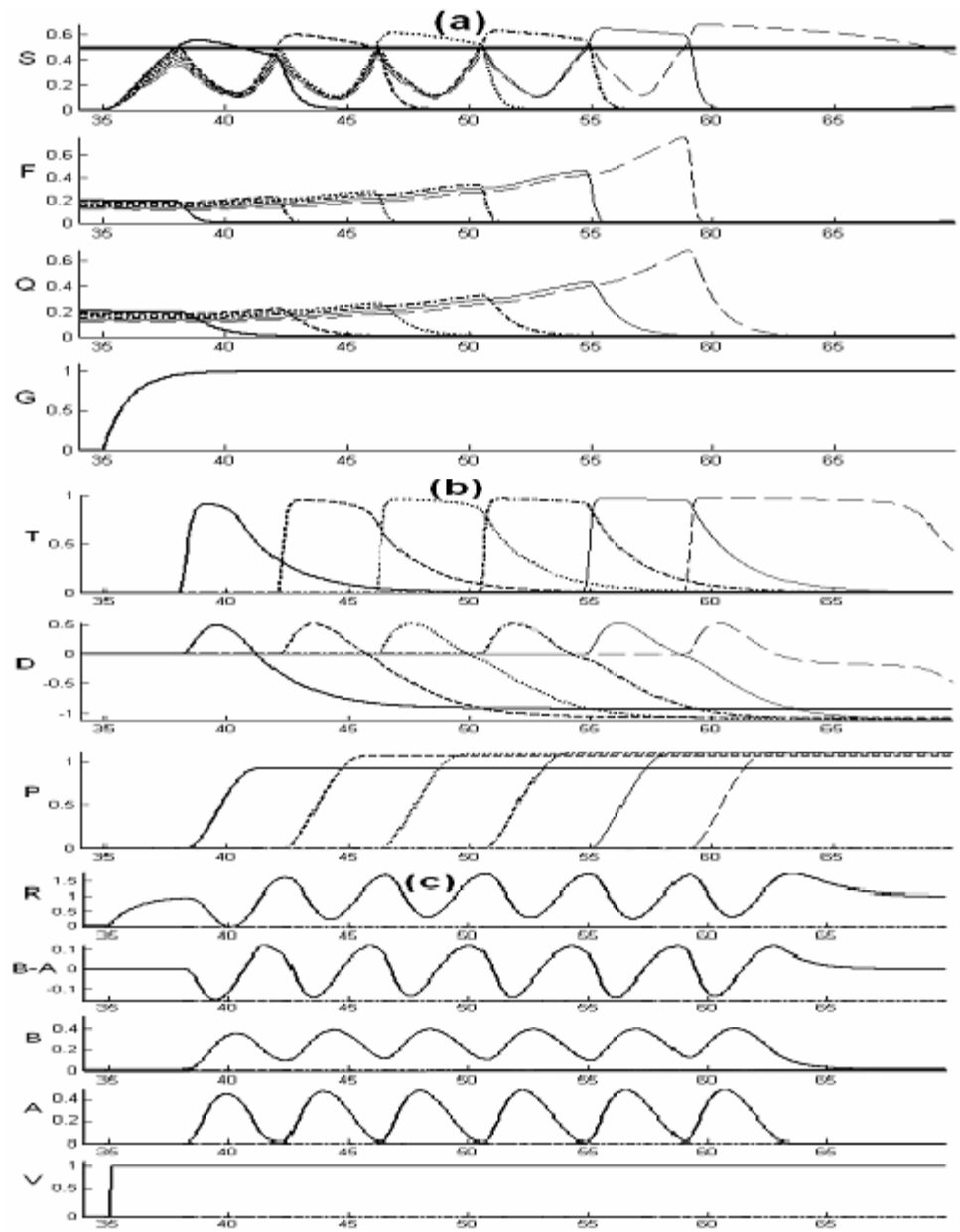


Figure 15. (a) Simulation of the motor working memory circuit when presented with a 6 item list with 5 simulation time unit IOIs. Inputs from cognitive working memory (*Y*) transfer an activation gradient to the motor plan storage network (*F*) and then are gated off by a volitional GO signal (*G*). The motor plan storage cells compete among themselves by sending inhibitory signals by way of inhibitory interneurons (*Q*). A rehearsal signal (*R*) allows read-out of the activity gradient by the plan selection layer (*S*), where a winner-take-all competition takes place. The winning cell activates a motor target (*T*) and sends feedback to inhibit its bottom-up motor plan (*F*). (b) Activities of the VITE trajectory generation circuit when presented with a 6 item list with 5 simulation time unit IOIs. Target position cell (*T*) is activated by a plan selection cell (*S*). An outflow representation of present position (*P*) is subtracted from the target position to compute a position difference (*D*) that represents the distance and direction the motor effectors must move. A volitional GO signal (*G*), which scales with desired rate of

movement (V), gates the position difference (DG) representation before it is integrated by the present position (P). Thus DG represents outflow velocity. This velocity signal is sent to cells A and B which time-average them at different rates. See Figure 14a. (c) Variable-rate rehearsal control circuit when presented with a 6 item list with 5 simulation time unit IOIs. The rehearsal signal (R) which gates the read-out of the motor plan gradient into the plan selection layer (S) receives volitional input that scales with performance rate (V_m), excitatory input from slow integrating velocity cells (B) and inhibitory input from fast velocity cells (A).

4.2.2. Sequential Rehearsal of Motor Plans. The control structures for rehearsal (Figures 1, 2, and 14) within motor working memory circuits help to realize the monitoring functions of dorsolateral prefrontal cortex (Petrides, 2005). In particular, a nonspecific rehearsal wave (R in Figures 2, 14a, and 15c, Appendix equations (21) and (14)) activates read-out of the motor plan activity gradient (Figures 2, 14a, and 15a, activities F_i in Appendix equation (9)) into a plan selection circuit (Figures 2, 14a, 14c, and 15a, activities S_i in Appendix equation (14)), whose recurrent on-center off-surround shunting network makes a winner-take-all (WTA) choice of the most highly activated item from the motor working memory. For proofs of how these WTA networks work, see Grossberg (1973) and reviews in Grossberg (1980, 1988).

The S_i whose activity rises first to a fixed threshold (.5) is selected for movement execution. Hanes and Schall (1996) have provided neurophysiological evidence that such a mechanism is how saccadic movement plans are chosen in FEF. The chosen activity S_i both elicits performance of its plan and self-inhibits its input source F_i ($100\left[S_i - .5\right]^+$ in Appendix equation (9)) to prevent item perseveration. How these movement commands are transformed into fluent, variable-rate, sequential actions under volitional control is now described.

4.3 Trajectory Generator

4.3.1. VITE Model of Motor Trajectory Control. A selected motor plan is transformed into a motor action via a trajectory generator (see Figures 2, 14b, 14c, and 15b). The trajectory generator that is used in the LIST PARSE model is a variant of the Vector Integration To Endpoint, or VITE, model. Since its introduction in Bullock & Grossberg (1988a), the model has been further developed to explain movement kinematics for planned arm movements, speech articulator movements, cursive handwriting, ball catching, and piano performance (Bullock, Bongers, Lankhorst & Beek, 1999; Bullock & Grossberg, 1988a, 1988b, 1989, 1991; Bullock, Cisek & Grossberg, 1998; Bullock, Grossberg & Guenther, 1993; Bullock, Grossberg & Mannes, 1993; Cisek, Grossberg & Bullock, 1998; Dessing, Peper, Bullock & Beek, 2005; Grossberg & Paine, 2000; Guenther, 1994, 1995; Guenther, Ghosh & Nieto-Castanon, 2003; Guenther, Hampson & Johnson, 1998; Jacobs & Bullock, 1998; Paine, Grossberg & Van Gemmert, 2004). VITE works as follows:

The chosen motor plan activates a Target Position Vector, or TPV (Figures 2, 14a, 14c, and 15b, T_i in Appendix equation (15)). The TPV represents a target configuration of motor effectors; that is, where the effectors are commanded to move. This target vector is compared with an outflow representation of the Present Position Vector, or PPV (Figures 2, 14b, and 14b, P_i in Appendix 1 equation (17)) which represents the present configuration of the motor effectors. The PPV is subtracted from the TPV to compute a Difference Vector, or DV (D_i in Figures 2, 14b, 14c, and 15b, Appendix equation (16)) which computes the direction and distance that the motor effectors need to move in order to realize the TPV. A volitional GO signal gates, or multiplies, the DV before it is integrated by the PPV; that is, the PPV integrates

the product (DV)(GO) (Figures 2, 14b, and 14c, $[D_i]^+ G$ in Appendix equation (17)). Because PPV integrates (DV)(GO), (DV)(GO) represents an outflow representation of outflow movement velocity (Figure 14b) which provides the basis for the velocity profile simulation (Figure 16b below).

Because read-out of a new TPV can occur without read-out of the volitional GO signal to perform it, a TPV can prime a DV before the volitional GO signal initiates movement. Increasing the GO signal increases movement speed. The PPV always approaches the TPV when the GO signal is positive, no matter how large the GO signal is chosen, because the GO signal multiplies the DV, which approaches zero as the PPV approaches the TPV.

4.3.2. Synergy, Synchrony, and Speed. A VITE circuit naturally embodies the Three S's of movement trajectory control: Synergy, Synchrony, and Speed (Bullock & Grossberg, 1988a). When the same GO signal multiplies a DV that controls a prescribed set of muscles (i.e., a synergy), those muscles contract in equal time (i.e., synchrony) and reach their respective TPV targets simultaneously, no matter how large the GO signal may be chosen (i.e., variable speed). Bullock & Grossberg (1988a) proved how such a trajectory controller can generate the bell-shaped velocity profiles that are ubiquitous in biological movement control, among other behavioral and neurobiological movement data.

4.3.3. Neurophysiological Data. Neurophysiological correlates of TPV, PPV, and DV have been reported in Brodmann's area 4 (primary motor cortex) and parietal area 5. These data are simulated and reviewed in Bullock, Cisek & Grossberg (1998). Evidence for GO signal properties have been reported for the case of arm movement control in the ventrolateral segment of the globus pallidus (GPi) or in the ansa lenticularis of the basal ganglia by Horak & Anderson (1984a, 1984b), among others. Brown, Bullock & Grossberg (1999, 2004) modeled how such basal ganglia gating may occur, notably how selective gating enables the brain to balance between reactive and planned movements, and regulates the learning of such planned movements, including the kind of data reported by Alexander & Crutcher (1990a, b), Alexander, Crutcher & DeLong (1990), Alexander, DeLong & Strick (1986), Hikosaka & Wurtz (1989), and Redgrave, Prescott & Gurney (1999).

4.4 Variable-Rate Volitional Control of Sequential Actions

4.4.1. Coordinating Variable-Rate Rehearsal and Performance. Volition controls variable-rate rehearsal and recall of action sequences. The stored items and their order in cognitive working memory are not known at the source of the volitional signal. Volitional signals are therefore *nonspecific*: a signal of equal size influences *all* of its target cells. The internal network structure translates the nonspecific signal into contextually appropriate movements.

Volitional signals are coordinated to enable fluent variable-speed performance. The model proposes that the same volitional signal, V_m, activates both the rehearsal wave activity R, which begins the read-out of motor plans from the motor working memory (Figure 14c, Appendix equation (21)), and the GO signal, G, which supports the execution of each motor command in the trajectory generator (Figures 14b and 14c, Appendix equations (17) and (18)). This hypothesis predicts that a basal ganglia volitional signal, V_m, projects in parallel as a rehearsal wave, R, to dorsolateral prefrontal cortex and as a GO signal, G, to cortical area 4 (Bullock, Cisek & Grossberg, 1998)). In particular, when V_m turns on, so does G. As noted in Section 4.2.1, onset of the G inhibits read-out of the cognitive working memory activities, Y_i, to the motor working memory activities, F_i. In this way, the volitionally triggered onset of a movement from the motor working memory is coordinated with the gating off of output signals from the

cognitive working memory to the motor working memory. The sequence to be performed by the motor working memory is now fixed while G remains on.

Larger values of V_m, and thus R and G, coordinate the rate at which items are read-out from motor working memory with the rate at which actions are performed by the VITE trajectory generator (Figure 15). As illustrated in Figures 2 and 14a, a larger G implies a faster integration rate within the motor working memory (term $(1+G)$ in Appendix equation (10)), faster selection and rehearsal of an action by the plan selection network (term $[R]^+$ in Appendix equation (14)), and faster performance by the trajectory generator (term G in Appendix equation (17)).

4.4.2. Anticipatory Selection of Actions: Distance and Speed Commands Are Not Enough. Volitional control in the model can also lead to anticipatory selection of the next motor plan in a sequence, and preparation for its execution prior to the successful completion of the preceding movement. As noted by Lashley (1951), in the absence of such anticipation, the performance of any novel sequence of movements would degenerate into a procession of disjointed individual movements separated in time by at least two transmission delays: one when a sensory cue of movement completion is sent to the brain, and the other when transmission of the next motor command is sent from the brain to the motor effectors.

What signals are used to control variable-rate anticipatory timing of sequential performance? The simplest cue, an estimated distance to target, while valuable, is insufficient for anticipatory timing because it cannot differentiate movement speed from movement size. In particular, the process that initiates the next movement must be triggered when the motor effecter is farther away from a target for a fast movement than a slow one, and the same distance to target could be part of the near completion of a large movement or the beginning of a small one.

A movement velocity-sensitive cue may help to differentiate movement speed from movement size. However, as noted in Section 4.3.1, (DV)(GO) represents the commanded movement velocity, and has a bell-shaped velocity profile as the DV is integrated through time until PPV equals TPV. Such a bell-shaped velocity profile cannot differentiate movement size, because it is close to zero both near the beginning and end of a movement. Moreover, movement sequences are often performed in a smooth way, without the abrupt starts and stops between individual movements that Lashley (1951) discussed. Thus, velocity may not approach zero between movements, except at the end of the entire movement sequence, thereby eliminating a near-zero velocity signal as an anticipatory trigger of the next movement.

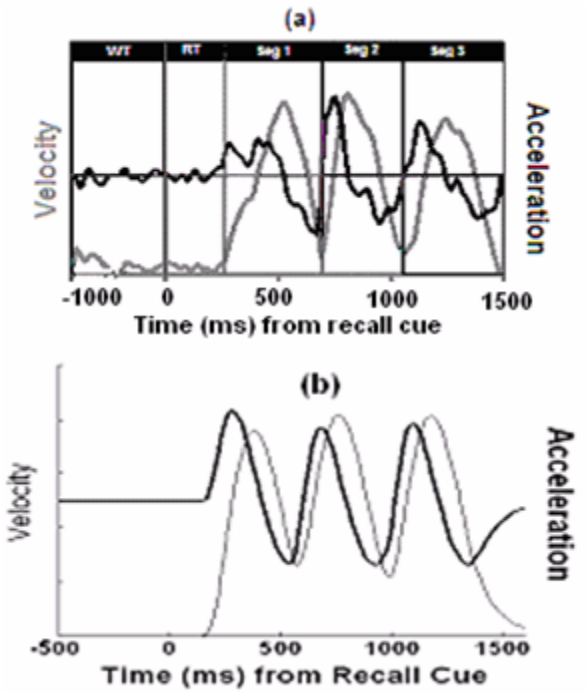


Figure 16. (a) Profile of movement kinematics (velocity and acceleration) from performance of a triangle during a monkey sequential copying task. (b) Internal estimates of velocity (*B*) and acceleration (*A* - *B*), based upon outflow velocity signals from the VITE trajectory generator, used by LIST PARSE model to construct the rehearsal signal (*R*). Note: in both graphs, the X-axis labels have been shifted so that 0 ms corresponds to the recall cue presentation time and the acceleration graphs have been shifted upward in both (a) and (b) to more easily portray the phase differences with the velocity graphs. (Adapted with permission from Averbeck et al., 2003a.)

4.4.3. Acceleration/Deceleration Estimates Influence Rehearsal Timing. Whereas distance and velocity cues may be insufficient to anticipate movement completion, movement acceleration and/or deceleration estimates are not. Given a typical bell-shaped velocity profile (Figure 16a), the acceleration function is positive for the first part of a movement, negative for the last part of a movement and, for smooth movements, tends to approach, but not equal, zero near the completion of a movement. This asymmetry between acceleration at movement initiation and completion provides information that the bell-shaped velocity profile may not. However, movement trajectory generators do not seem to explicitly compute accelerations and decelerations. Rather, they explicitly compute difference vectors and outflow velocity estimates (Figure 14b). How can such a movement controller naturally compute estimates of movement acceleration and deceleration?

LIST PARSE proposes that such estimates are derived from the bell-shaped velocity profile in the following way: In addition to sending signals to the PPV stage, where they are integrated into outflow present position commands, the (DV)(GO) stage also sends signals to two parallel cell populations that time-average the velocity signals at different rates. In other words, the (DV)(GO) stage sends signals to three cell populations that time-average it at different rates, with the PPV stage having the smallest decay rate to act as a movement integrator.

To compute movement acceleration, *fast cells* time-average the velocity signal more quickly (Figure 15c, activity *A* in Appendix equation (20)) than *slow cells* (Figure 15c, activity *B* in

Appendix equation (21)). During the beginning of a movement, the bell-shaped velocity curve increases. Hence activity A is greater than activity B, because fast cells better track the increasing velocity signal, whereas slow cells still remember smaller velocity signals. At the end of a movement, the bell-shaped velocity curve decreases. Here, activity A is less than activity B, because fast cells better track the smaller velocities, whereas slow cells still remember the larger velocities from when the bell-shaped velocity profile was still increasing.

The model proposes that activity B excites the rehearsal cell activity R, whereas activity A inhibits it (Figure 15c, Appendix equation (21)). At the beginning of movement, this difference signal ($B - A$) is negative, thereby shutting off the rehearsal gating signal, R, and preventing premature read-out of the next motor item. At the end of movement, ($B - A$) is positive and can open the rehearsal gate to initiate the next movement before the present movement is over. Since ($B - A$) is derived from the bell-shaped velocity profile, it is sensitive to the rate of movement and can control seamless performance of a movement sequence at variable speeds.

Anticipatory initiation of motor commands for piano key presses and visually guided catching or hitting have also been modeled using an explicit computation of an estimated time-to-contact signal (e.g., Bullock et al., 1999; Dessing, Caljouw, Peper & Beek, 2004; Dessing et al., 2005; Jacobs & Bullock, 1998; Lee, 1976). While remaining agnostic about the use of time-to-contact, LIST PARSE shows how velocity outflow signals can directly control rate-sensitive, anticipatory control of sequential movements.

5. Data Simulations and Methods. Model processes of the above type are, we would argue, needed to simulate real-time sequential behaviors. Even the simplified current LIST PARSE versions of these processes embody a neural model of significant complexity. In order to maximize clarity and understanding, a modular approach to the simulations was adopted. Although system integration is needed to understanding the entire process, simulations were carried out, wherever possible, by the model processes that are rate-limiting in explaining particular data, rather than running all data through the entire model.

5.1 Serial Recall Simulations. The item presentation protocol for the simulations in all tasks is defined in Appendix Section A.2. Serial recall experiments were chosen for simulation which minimize or eliminate covert rehearsals and the associated confounds that such rehearsals entail, including only experiments with very short IOIs, and the presence of distracter tasks (e.g., Cowan, 1999; Murdock, 1961), or explicit instructions not to rehearse (e.g., Cowan, 1999). Consequently, no rehearsals are simulated in these serial recall tasks.

Model parameters relevant to order error production (variables b, e, and F in Appendix equations (2) and (3), and the variance of Gaussian noise, .015; see Table 3 in Section 5.7) were selected based upon an extensive grid search to provide a best fit for the performance curves in Experiment 1 of Henson et al. (1996, Figure 6). Given the small proportion of omissions in span and sub-span ISR tasks, the output threshold (.165) relevant for omission, or item, errors was selected to provide a best fit for the performance curves in Experiment 1 (fast condition) of Tan & Ward (2000, Figure 8). These parameters were then fixed in all ISR, delayed serial recall, and IFR simulations (Figures 6, 7, 8, and 9).

Item errors make up a tiny percentage of errors for span and sub-span lists. They occur in the model when the activity of an item in cognitive working memory falls below the output threshold for being transferred to motor working memory, or activity of an item in motor working memory falls below the output threshold for being selected for performance (Figure 10b). A modular approach to simulating these data is now summarized.

The majority of errors in span or sub-span serial recall (Figures 6, 7, and 8) are order errors, which are attributed to noise that distorts the relative activities of items stored in cognitive working memory and/or motor working memory. This stored pattern, which is ideally a primacy gradient, emerges from the neural interactions between layer 4 and 6 cells (Appendix equations (2) and (3)). Due to competitive interactions, an item's activity and the relative activity differences between items are influenced by the number of items being stored and the length of time they are stored. The susceptibility of an item's temporal order to noise fluctuations is thus also sensitive to these factors.

The effects of noise on the stored cognitive working memory gradient were simplified to reduce simulation time and simplify data fitting: Order errors in the serial recall simulations (Figures 6a, 7, 8a, and 8c) were assumed to occur due to Gaussian noise. Rather than having noise operate continuously in time, zero mean Gaussian noise with variance .015 was added to layer 6 activities Y_i (Appendix equation (3)) at a single time, five simulation time units after the time of the recall cue. Item omissions are counted if the corresponding activity Y_i, after noise acts, does not exceed the output threshold (.165).

These simplifying assumptions take advantage of the fact that the largest influence of noise and thresholding occurs just before the transfer of information between working memory stores. The error patterns generated in this way are similar to those produced by continuous noise because activities in cognitive working memory continuously compress throughout time, becoming more susceptible to order errors the longer multiple items are simultaneously held. The self-normalizing dynamics in motor working memory make additional item and order errors progressively less likely as items are selected and suppressed, because their activity is redistributed to the remaining items, thereby expanding their activity range.

The simulations that generated Figures 6, 7, and 8 computed the noise-free cognitive working memory activities until noise acts. Then 100,000 independent noise addition trials were carried out. In each such trial, the largest item was selected, compared to the output threshold to see if it is recalled at all, and set to zero to prevent perseveration. Then the process was iterated for the next largest item. Running 100,000 full network simulations in which stochastic processes continuously operated at multiple levels in the full network would be computationally infeasible, particularly during parameter fitting when hundreds of different parameter starting points are examined.

5.2. Modality Effects. Modality-related recency effects include the advantage in recall for the last one or two items for audibly presented lists as compared with visually presented lists. Page & Norris (1998) suggested that such recency effects may occur because, during auditory list presentation, the last presented item is recalled from a reliable independent processing stage, or "transient sensory store," such as the pre-categorical acoustic store of Crowder & Morton (1968). The last item is recalled reliably, as long as no additional auditory information displaces it. The lack of this item's interference with other items in working memory can improve their performance as well, particularly the last few items which are most likely to interchange with this last item. Watkins & Watkins (1980) provided evidence that such a reliable limited capacity store may persist for more than four seconds in the absence of additional phonological information. This assumption also provides a possible explanation for the *suffix effect* (Crowder, 1978), in which the addition of a single irrelevant auditory item greatly attenuates modality differences.

In LIST PARSE, in response to audibly presented lists that use written recall, the last presented item is presumed to be recalled nearly perfectly, due to being stored in a separate

echoic memory buffer, and is not present in the competitive working memory. Read-out from echoic memory is performed algorithmically in the model. The echoic processing level from which the last item is read-out into motor production is hypothesized to be part of the auditory perception/production loop that is learned through babbling and imitation of other speakers and that seems to have a homologous design to the one used to learn reaching movements through babbling and imitation (Bullock, Grossberg & Guenther, 1993; Cohen, Grossberg & Stork, 1988; Grossberg, 2003b; Guenther, 1995, 2006; Guenther, Hampson & Johnson, 1998).

5.3. Primacy, Bowing and the Serial Position Performance Curve. Most ordinal models of sequence representation (e.g., Farrell & Lewandowsky, 2002; Page & Norris, 1998) *assume* a primacy gradient and then use it to explain cognitive data. These models attribute decreased order error rates for the first and last items in a list *entirely* due to end effects (Page & Norris, 1998). That is, order errors are explained as a result of noise that causes an item to become more or less activated than another item, leading to its selection earlier or later than is appropriate. Items with the largest and smallest activities have fewer items with similar activation levels with which to interchange. The one-item recency portions observed in the sub-span ISR protocols in Figure 6 and in the peaks (circles) of Figure 7 could be due to such end effects.

Figure 6a demonstrates simulations of the serial position curve data of Cowan et al. (1999) in Figure 6b. These experiments used quickly presented (two per second, 500ms IOI) auditory lists of 3-8 items and a visual distracter task to prevent rehearsal. Given that these were auditory lists that required written recall, the last item in the list was assumed to be recalled nearly perfectly and kept separate from the competitive working memory interactions and the effects of noise.

Performance was nearly perfect for shorter lists (3 and 4 item lists) in both the data and model simulations, and were omitted for clarity of representation in Figure 6b. LIST PARSE imulations closely match data for all list lengths in the attended speech condition of the Cowan et al. (1999) study, with the most significant departure being that subject performance of the first item of every list length tested was perfect (Figure 6b). These fits are impressive because the ISR task that the parameters were selected to fit (Experiment 1 of Henson et al., 1996, Figure 6) only studied one list length (six visually presented items). Despite having an identical IOI of 500ms, performance for the first item at this list length in that task was significantly below 100 percent. One possible difference between the two tasks was the presence of a two second retention interval in Cowan et al. (1999), which could have been used to covertly rehearse the first couple of items immediately before recall.

A purely end effect account of recency gradients is contradicted by the existence of recall protocols in which a recency portion of more than one or two items appear (e.g., verbal ISR: Cowan et al., 1999; Klein et al., 2005; spatial ISR: Jones, Farrand, Stuart & Morris, 1995; IFR: Tan & Ward, 2000). How can such extended recency gradients be explained? In contrast to other ordinal models (e.g., Page & Norris, 1998), the working memory gradients in LIST PARSE are emergent properties of network interactions as a list is presented to the network in real time; they are not just mathematically assumed properties. These emergent properties predict that end effects are only part of the explanation for the recency portion of the ISR, or IFR, serial position curves. In LIST PARSE, cognitive working memory dynamics, even in the absence of rehearsal, may yield extended recency gradients as emergent properties (Figure 12c) that depend upon list length, rehearsal schedule, volitional and attention.

Such bowed activity gradients do not, however, bowing of the same shape in the *total* error serial position cures during serial recall tasks. Bowed gradients naturally lead to a higher probability of recall for the final items in extended lists, and thus an extended bow in *item* errors.

However, as explained below, LIST PARSE predicts that recency gradient items would tend to be recalled earlier in the list than is appropriate in the ISR task, resulting in pronounced *order* errors for these recency items. For super-span ISR tasks, a bowed activation gradient predicts significant recency portions of the *item* serial position curve, with less prominent recency effects in the *order* and *total* (both item and order errors) serial position performance curves. All these properties of error distributions are, in fact, observed in super-span ISR tasks (Drewnowski, 1980; Drewnowski & Murdock, 1980; Watkins & Watkins, 1980; Klein et al., 2005; for a review: Henson, 1998b).

5.4. Error Type Distributions. The ISR task from Experiment 1 in Henson et al. (1996) used six item, visually presented lists of letters with 500ms IOIs and written recall. A simulation of the error distribution from these data is presented in Figure 7. This simulation was created using the same basic procedure as for all of the ISR data, with five simulation unit IOIs, and no covert rehearsal. Because this experiment used visually presented lists, no (modality-related) advantage for the final list item was assumed. These data determined the free parameters in Appendix 1 equations (2) and (3), as well as the variance of Gaussian noise. In Figure 7, each solid line plots the probability that an item that is presented in a given serial position will be recalled in serial positions 1-6. The circles represent the probability of each item being recalled in the correct serial position (i.e. the serial position performance curve). Item activities representing neighboring sequential recall positions differ less than the activities of more distant items (see Figure 12b). Gaussian noise thus tends to preferentially increase interchanges of items in neighboring serial positions, as shown in Figure 7. If an item from later in the list erroneously acquires the largest activity and is selected, the remaining items will still be biased toward a primacy gradient, and thus the most likely item to be selected will be the skipped item. Thus, the model naturally produces fill-in effects (Norris, Page, & Baddeley, 1994).

5.5. List Length Effects. Increasing the number of items to be stored has a negative influence upon the ability to remember the entire list. Crannell & Parrish (1957) used visually presented lists to show that increasing list length produces an S-shaped curve in the ability to correctly recall entire lists. Figure 8a summarizes a simulation of data from Crannell & Parrish (1957, circles) and Baddeley & Hitch (1975, squares). These simulations used the same procedure as for the Henson et al. (1996) data in Figure 7, with the addition that simulations were performed for lists of 5 - 10 items. Since correct performance of the entire list is at issue in the Crannell & Parrish (1957) and Baddeley & Hitch (1975) data, only simulated recall trials that produced recall *of every item in the list in the correct serial position* were counted as correct.

LIST PARSE predicts that longer lists result in increased list errors for several reasons: First, when more items are stored, normalizing competitive dynamics in layer 4 decreases both the absolute and relative activities of stored items. These properties lead to increased omissions in layer 6 as list length increases. In addition, again due to competitive normalization, additional stored items provide more opportunities for interchanges, and thus order errors, due to noise. Another factor is that the time needed to present the additional items, while maintaining the same IOIs, results in longer durations that the first items must be stored. Their activities therefore compress together more (Figure 12a), and hereby increase interchange probabilities. Finally, lists whose length is well above span begin to produce bowed gradients (Figures 12c), so noise is needed for the final items to become least active and recalled last.

5.6. Temporal Limitations Upon Recall. A delay prior to recall, during which rehearsal is prevented by distracters, dramatically reduces correct recall. Such delay effects were reported by Murdock (1960, 1961) in a serial recall task using sub-span length lists of letters and words. Data

from Murdock (1961) are summarized in Figure 8b, and model simulations in Figures 8c. These simulations used the same procedure as for the Crannell & Parrish (1957) and Baddeley & Hitch (1975) data in Figure 8a, with the caveats that simulations were performed only for lists of 3 items, and performance was assessed after unrehearsed retention intervals of 1-24 seconds, again with the convention that 1 simulation time unit equals 100ms. The percentage of trials at each retention interval that produced accurate recall of the entire list was then plotted (Figure 8c).

After approximately 9 seconds in both the data and simulations, performance plateaus at approximately 20% (see Figure 8b). Although Murdock (1960) did not present separate results for item and order errors in this task, LIST PARSE predicts that, for sub-span lists (3 items in this task), unrehearsed retention intervals particularly harm order retention, again due to the increased time that item activities have to compress (Figure 12a). The observed performance plateaus occur in the model because the item activities become so similar that noise dominates recall order. Pure chance would cause a plateau performance level of 16.7% (1 out of 6 possible orderings of 3 items). Said in another way, the model retains item information better than order information for sub-span lists, consistent with the 1 item high performance plateau (Figure 8b).

5.7. A Unified View of Free and Serial Recall: Same Working Memory, Different Selection Strategies. LIST PARSE proposes that the working memory representations underlying serial recall and free recall are the same, and thus that a unified account of these widely studied, but typically disparate, working memory protocols is possible. This proposal follows naturally from the Grossberg (1978a, 1978b) prediction that all working memories share similar designs in order to permit stable learning of list chunks (see Sections 2.4 and 2.5). However, differences in the serial position curves from the two tasks seem to contradict such a proposal: Most serial position curves in serial recall exhibit a primacy gradient with perhaps one-item recency, whereas free recall tasks typically exhibit more rounded serial position curves with pronounced recency effects.

We would argue that several factors create these different data properties: Before even discussing different mechanisms, one needs to acknowledge that curves in the two tasks usually derive from different IOIs, numbers of items, and error definitions, with serial recall including both item and order errors. Even when the protocols are identical, rehearsals are rarely controlled, so different rehearsal strategies could enable the stored working memory activity patterns to become differentiated from one another. Rehearsal strategies can have a significant effect when the order of rehearsal is constrained by different paradigm requirements, such as the requirement of recalling items in their correct order or not. In particular, free recall allows the use of transient sensory stores to pick off the last item(s) at the beginning of the recall period. Few “free” recall tasks require subjects to begin at the list beginning to prevent this strategy, and few serial recall tasks allow subjects to start recall at the end of the list to possibly utilize such stores.

Even in the absence of rehearsal, the same cognitive working memory dynamics that produce primacy gradients in sub-span and span length serial recall cases (Figure 12b) can produce bowed gradients with extended recency portions for super-span lists (Figure 12c). When the number of items and IOIs are equated in ISR and IFR tasks, and fair comparisons of errors are used, differences in serial performance profiles are greatly attenuated but still exist (e.g. Klein et al., 2005). What is the cause of these residual differences?

As noted above, rehearsal differences may be one such factor. During free recall, rehearsal strategies can differ between free recall trials and/or subjects, and rehearsal differences can alter free recall performance (Brodie & Prytulak, 1975; Kahana & Howard, 2005; Rundus, 1971; Tan

& Ward, 2000). For sub-span lists or lists with large inter-stimulus intervals (e.g., the slow condition of Tan & Ward, 2000), participants seem to mimic sequential (ISR) rehearsal strategies, which is easily explained if the same working memory representations are at work. In particular, many subjects rehearse, starting from the beginning of the list, as many items as possible during each rehearsal period in their presented order. However, for lists where performance is not nearly perfect, such as long lists, or lists with short inter-stimulus intervals, such a strategy would enable rehearsal of only the first item or two during every rehearsal period (Tan & Ward, 2000).

The LIST PARSE model demonstrates how differences in subject strategies during rehearsal and recall in the two tasks can explain differences in the data, even if both tasks use similar working memory representations. An explicitly modeled rehearsal procedure was used to simulate free recall data, as discussed in Section 5.8 and summarized in Table 3.

Table 3. Simulation procedures for serial and free recall tasks.

	Serial Recall	**Free Recall**
IOI	500ms (5 simulation time units: s.t.u.)	1250ms (12.5 simulation s.t.u.)
Item Presentations	1 s.t.u. duration; .1 magnitude	1 s.t.u. duration; .1 magnitude
Cognitive WM	$b = .2$; $e = .05$; $F = 1.25$	$b = .2$; $e = .05$; $F = 1.25$
Rehearsal Protocol	N/A because of the short ISI. Optimal strategy: (1) Cognitive WM activity is reset by reducing V_c (increasing F) after items are loaded into motor WM. (2) Items are re-presented to cognitive WM as they are rehearsed.	At each IOI midpoint: (1) Uniform random variable in [0, 1]. If activity is greater than expected number of rehearsals (Figure 7b), no rehearsal and re-presentation. Otherwise, execute one item rehearsal. Selection of item for re-presentation: (2) Uniform random variable in [0, .2]. Sum with stored cognitive WM activities. (3) Re-present item (excluding the item presented in current IOI) with maximum sum above selection threshold (.5).
Recall Protocol	(1) Record activity 5 s.t.u. after task. (2) Add Gaussian Noise to each activity in (1) (mean = 0; var = .23). (3) Output threshold at .165. (4) Ordered by decreasing activity. (5) Repeat 2-4 for 100,000 trials.	(1) Record activity 5 s.t.u. after task. (2) Add Gaussian Noise to each activity in (1) (mean = 0; var = .23). (3) Output threshold at .165. (4) Order irrelevant. (5) Repeat 2-4 for 100,000 trials. (6) Repeat 1-5 for 250 rehearsal orders.

In this regard, Dalezman (1976) has found that recall order can be a large source of the differences in serial and free recall performance. In this study, "free" recall performance was tested, but with an experimental condition for which subjects begin recall with the beginning of the list, rather than the more typical strategy of initially picking off the last few items in the list. Dalezman found that recency was greatly attenuated or eliminated when recall proceeded from the beginning of the list, and the typical enhanced recency effects in IFR were seen only when recall began from the end of list.

Also consistent with this view, Cowan, Saults, Elliott & Moreno (2002) used a protocol in which 9 item lists were presented for serial recall and then cued to begin recall with the 1st, 4th or 7th item. In the 4th and 7th item cases, subjects were required to continue reciting the list in the presented order by wrapping around to the list beginning. Performance for the last three presented items in the list was dramatically improved when they were recalled first. In fact, the last three presented items were recalled more reliably than the middle three or first three presented items in the conditions where they were the first recalled items.

5.8. Simplifying Assumptions and Simulation Protocol. In model simulations, presentations of new items in the IFR simulations occurred every 12.5 simulation time units, corresponding to 1.25 second IOIs (1 simulation time unit equals 100 ms) used in the fast condition of experiment 1 in Tan & Ward (2000). A model rehearsal consisted of a re-presentation of the rehearsed item (rehearsal selection procedure discussed below) with the same stimulus duration (one simulation time unit) and magnitude (.1) as an ordinary stimulus input (I_i), at an interval evenly spaced between two new item presentations (i.e., 6.25 simulation time units after the most recent item presentation onset).

The model assumes two differences in the manner in which rehearsal and item storage takes place during serial recall vs. super-span, short-IOI IFR tasks. These differences are summarized in Table 3:

Serial Recall Storage. Order information is best preserved when relative item activities are maximally differentiated. The cognitive working memory exhibits compressive competitive interactions and normalization. Maximal differentiation thus occurs when items are only stored in memory together for a limited time and are not all highly activated. The optimal rehearsal strategy is to remove items in cognitive working memory prior to rehearsal, so that items rehearsed from motor working memory are treated as a new list presented at more recent rehearsal times. Clearing cognitive memory can be realized by transiently reducing the V_c signal, thus increasing F (Appendix equation (2)).

Immediate Free Recall Storage. Item information is best preserved when items are maximally activated. Thus, eliminating items from cognitive working memory is now maladaptive. Letting re-presentation of an item during rehearsal add to the still active items in cognitive working memory is advantageous, and is assumed to occur in our free recall simulations.

Serial Recall Rehearsal. Preserving serial order is essential, and any other performance criterion is secondary. The model lets the motor working memory representations compete in the selection circuit as they rise to the output threshold.

Immediate Free Recall Rehearsal. Rehearsing as many items as possible in the allotted time is the goal. Rehearsal is speeded by adding positive uniform noise to the representations entering the selection circuit. The maximally activated, supra-threshold item is rehearsed. Thus, free recall rehearsal is probabilistic, but biased by the level of activation of each item at the time of recall. The rehearsal algorithm is discussed in more detail in Appendix Section A.2.

This rehearsal procedure predicts which single item will be rehearsed during any given rehearsal period (first recall probability), which is sufficient given the short IOIs of the task. LIST PARSE makes no claims about the ordering of multiple item rehearsals, or order of recall beyond the first item. The recall of one item undoubtedly biases which item will be recalled next based in part upon learned inter-item associations.

5.9. Self-normalizing Dynamics, Load Effects, and Rehearsal. Items that have higher activities when they enter the motor working memory have a proportionate advantage in

exceeding the output threshold quickly and being selected. Load effects can be understood if selection takes longer when more activity has to be added to exceed threshold. This follows because item activities are partially normalized, so that the average activity of items is less when more items are stored, and thus have farther to go to exceed threshold. The latency of the first item selection is greater with a larger working memory load, and thus there is a higher probability of "no rehearsals" within the fixed ISIs near the list end (Figure 9b). Load latency effects that are observed in sequential tasks (e.g. Sternberg, Monsell, Knoll & Wright, 1978) were qualitatively explained by self-normalizing working memory dynamics in Grossberg (1978a) and quantitatively simulated in Boardman & Bullock (1991) and Rhodes et al. (2004).

The variable rehearsal order in IFR simulations creates wide variations in the resulting activities of individual items in cognitive working memory at the time of recall from trial to trial. To characterize the distribution of possible activity gradients resulting from different rehearsal strategies, 250 separate trials were conducted with the stochastic selection processes guiding rehearsal throughout list presentation. In contrast, no rehearsal was assumed in the serial recall simulations, so this additional "rehearsal distribution" step was unnecessary (see Table 3). The activity gradients in both tasks were then distorted by noise to calculate error responses. Each activity set received 100,000 independent noise addition trials. If an item's distorted activation failed to exceed the omission threshold (.165) in either task, it was not recalled and constituted an item error. The super-threshold serial recall items were selected in order of decreasing activity and compared to their presented order to assess order errors.

5.10. Bowing and the Serial Position Performance Curve. As noted in Section 3.9, Figure 9 summarizes model simulations of an IFR experiment of Tan & Ward (2000). The average activity gradient in Figure 9a differs from the serial performance curve in Figure 9c, which summarizes the percentage of trials on which an item was recalled. Figure 9c has a much more prominent recency portion. Why is this? In Figure 9a, the final items do not have dramatically more activity on average than other parts of the list. However, they consistently have activities that exceed the output threshold when recall begins. The small lag between the presentation of the final items in the list and the onset of recall allows very little time for their activations to decay and fall below threshold, even though these items rarely have a chance before the beginning of recall to be rehearsed multiple times. Strong recency in Figure 9c results.

Figures 9d and 9e illustrate that, the later an item is rehearsed, the less time it has to decay and the more likely it is to be recalled. Item rehearsal boosts its activation and increases the likelihood that, in the time from the rehearsal to recall, the combination of leakage and normalizing dynamics do not drop its activity below the omission threshold.

In Figure 9f, the increased number of rehearsals of early list items and the tendency of these items to be rehearsed later in the list reinforces a primacy gradient in the serial performance curve.

These network dynamics help illustrate why delays after list presentation preferentially attenuate the recency portion of the serial position curve (Figure 9g); cf., Glanzer (1972), Glanzer & Cunitz (1966), and Postman & Phillips (1965). The dotted and solid curves represent model simulations which differ only in the addition (in the dotted case) of a 12.5 second delay period after list presentation before the application of omission thresholding. Strong recency effects in immediate serial recall (i.e., the most recently presented items being recalled on most trials) are not due to the final items being very strongly active. Instead, these items are just active enough, due to a lack of time to decay, to exceed threshold on almost every trial. By including a period—without rehearsal—during which items can decay and self-normalize after list

presentation, these "just active enough" items are highly susceptible to falling below threshold. The primacy portions of the curve are less susceptible to this post-list delay effect because, when they are recalled, it is due largely to their being rehearsed multiple times throughout the list and thus having activations that greatly exceed threshold.

The dashed-dotted curve in Figure 9g illustrates the case of continuous-distracter free recall, wherein there is not only a long post-list delay period with no rehearsal (identical to the simulation of delayed free recall above) but distracter tasks are also rapidly interleaved with item presentations. In this demonstration, the 1.15 second distracter periods following the 100ms item presentations not only prevent any rehearsal from occurring but are also assumed to diminish volitional gain control of working memory, a sort of divided attention due to rapid task switching between presentations and distracters. This is modeled as a switch of parameter F in Appendix Equation (2) from a value of 1.25 to 2 during these 1.15 second periods following list item presentations. The result is a weaker strength of encoding for primacy items in the list than is seen in the delayed free recall case, where rehearsal can proceed unfettered and no inter-list distracters interfere with the strength of encoding for these items. Recency portions benefit from this reduced representation of earlier list items, due to less interference, and the result is stronger recency in continuous-distracter than delayed free recall. Were LTM-STM interactions included in these simulations, this effect would only be magnified.

A number of dissociations between immediate free recall and the continuous-distracter free recall paradigm have been described by Davelaar et al. (2005). The current article does not attempt to explain them all due to its focus on working memory dynamics and subject strategies.

5.11. Sequential Movement Neurophysiology. LIST PARSE models a neural system that integrates cognitive, motor, and volitional dynamics (Figures 1 and 2). The motor and volitional dynamics are needed to recall, or perform, even the most basic sequential tasks. These circuits were omitted from the simulations above because they are not rate-limiting in explaining the targeted cognitive data. However, these circuits were tested in simulations of data about sequential arm movements. The fact that verbal working memory data in humans and motor working memory data in monkeys can both be simulated by the same model provides further evidence for the hypothesis that all working memories in the brain share a similar design.

As noted in Section 3.2, neurophysiological recordings from the dorsolateral prefrontal cortex of monkeys have demonstrated that a parallel activity gradient controls planned sequential movements, with larger total activity corresponding to earlier sequential performance (Averbeck et al., 2002, 2003a, 2003b). The Averbeck et al. (2002) data in Figure 10a are simulated by cell responses, F_i, in the motor working memory in dorsolateral prefrontal cortex. The motor plan field cells in F_i interact within a recurrent shunting on-center off-surround network, or recurrent competitive field (Grossberg, 1973, 1980). Such shunting competitive dynamics have, for example, been used by Chey, Grossberg, and Mingolla (1997), Cisek (2006), and Usher & McClelland (2001) to describe the time course of perceptual and motor choice. The shunting competitive interaction leads to the property of self-normalization; i.e., the total activation across the network is approximately the same over time. Thus, reduced activation of one cell population leads to an increase in activation of other active cells in the network.

The cell activities, F_i, were produced by presenting input sequences of length 3, 4, and 5 with five simulation time unit IOIs, identical to the methodology used in the serial recall simulations (Figures 6, 7, and 8). The motor working memory activities, F_i, were held at zero until the 25th simulation time unit across all list lengths, at which point the activities in cognitive working memory were allowed to flow into the motor plan field. This procedure mimics buffer loading

from cognitive working memory into motor working memory. The activity pattern that is stored in motor working memory at the time of the recall cue (at the 35th simulation time unit) is the same whether this algorithmic buffer loading procedure is used or items are continuously loaded from cognitive working memory.

As task performance proceeds through each phase, as discussed in Section 3.2, the following response properties of F_i arise: The primacy gradient activities Y_i in cognitive working memory are transferred to motor working memory activities F_i prior to initiation of the first movement. This primacy gradient is contrast-enhanced to favor the most active item prior to performance of that item. Contrast-enhancement is due to increasing a volitional GO signal, combined with a visual feedback signal generated by the eye movement that precedes the arm movement (term $\sum_{k \neq i} E_k$ in Appendix Equation (10)). This hypothesis is consistent with the monkeys' reported eye movement patterns (Averbeck et al., 2003a, p. 132): "The monkeys generally made a saccade to the template after it appeared [in the representative figure, the monkey foveated the first segment of the eventual movement sequence]. They then made a saccade back to the drawing area as drawing began, followed by a sequence of smaller saccades falling near the advancing copy trajectory as this progressed."

Next, the F_i activities drive the plan selection activities, S_i, until the maximally activated plan exceeded a fixed threshold and is selected (Appendix equation (14)). This selected plan is forwarded to a motor target position representation, T_i, of a VITE trajectory generator (Appendix equation (15)). As this happens, feedback inhibition (S_i to F_i) suppresses the chosen representation in F_i, and thereby prevents perseveration of the same motor command (term $100\left[S_i - .5\right]^+$ of Appendix Equation (9)). The self-normalizing dynamics of the motor plan activities F_i enhance activities of the other active motor plans and facilitates their performance.

The data hint that the primacy gradient may be re-established when all sequence items have been selected, although this information was not collected during the experiment: See the right-most increasing data traces in Figure 10. The self-normalizing dynamics of the cells in F_i would result in re-establishment and re-performance of the list unless the volitional signal, V_m, to continue rehearsal is turned off (see Figures 2 and 14).

Different parameters were used to fit human cognitive data and monkey sensory-motor data. The shape of the motor working memory primacy gradient prior to initiation of the monkey movement sequence (Figure 10) differs from that of the cognitive working memory used to fit the human cognitive data (Figures 12). At least three possible reasons for this difference are the different species of subjects: macaque monkeys vs. humans; the different movement tasks: arm control during copying of a concurrently presented figure versus visual or verbal recall of sequentially presented items; and different levels of learning: the monkeys in the copying task knew which shape would need to be copied, so performance of the sequence was well-learned and practiced, whereas the cognitive tasks involved novel sequences of familiar items. Even assuming shared brain mechanisms between tasks and across species, any of these factors could account for something as simple as different levels of attention, which could distort the gradient in cognitive and motor working memory. While remaining agnostic as to the reasons for the difference in gradient shape, the variables of the cognitive working memory which determine this gradient shape (b = 2, e =. 2 and F = 15 Appendix equations (2)-(3)) were hand-chosen to fit the gradient in F_i at the marked time (Figure 10b) for the triangle in the Averbeck et al. (2002)

data. These parameters were then used to produce the simulations for the three (trapezoid), four (inverted triangle) and five (square) item sequences in Figure 10b.

5.12. Velocity and Acceleration Approximations. As noted in Section 4.4.3, LIST PARSE proposes that internal estimates of velocity (DV)(GO) and deceleration (*B* - *A*) are used to anticipatorily time rehearsal of even novel sequences of well-learned movements. The pattern of these internal representations of velocity and acceleration (Figure 16b) closely match the observed kinematics recorded from monkeys performing the Averbeck et al. (2002) copying task (Figure 16a). Note that the acceleration signal that is used to control timed rehearsal of sequential items is really a *de*celeration (1 – acceleration), whereas the behaviorally observed *ac*celeration is plotted in Figure 16b.

5.13. Temporal Grouping Effects. Temporal grouping effects (Ryan, 1969) occur when items enter cognitive working memory at nearly the same time; that is, are temporally grouped by having short IOIs. Because their activities are integrated in cognitive working memory for similar durations, they have similar activities. Inserting a long pause, or large IOI, in the middle of a longer sequence results in significantly different integration times for working memory items before vs. after the pause, and thus large differences in the activities of these two groups of items. As a result, order errors among temporally grouped items are more likely than errors across temporal groups, and end effects, and thus bows, occur within these temporal groups. These are well-established effects of temporal grouping (e.g., Ryan, 1969). They are illustrated by the simulation of relative *order error proportions* produced by LIST PARSE when a 6 item list with 500 ms IOIs has an additional pause of 200 ms inserted between the presentations of the third and fourth item (Figure 17a). All of the other simulation protocols (e.g., stimulus duration, magnitude of stimulus, cognitive working memory variables, noise variance, etc.) were identical to those used in the simulations of the Henson et al. (1996) data in Figure 7.

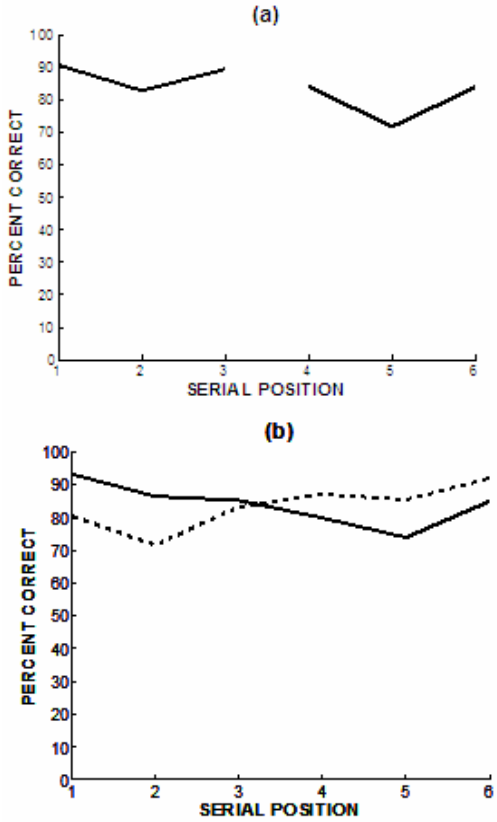


Figure 17. Simulation of the proportion of order errors vs. serial position for 6 item lists with: (a) an extended pause (IOI equal to 7 time units) between the third and fourth items, and IOIs of 5 simulation time units between other items, and (b) IOIs of 5 (solid) and 10 (dashed) simulation time units between all items.

Variations in experimental protocols (e.g., IOI) can influence the distribution of errors even when rehearsal is prevented. However, few studies have used long IOIs while simultaneously employing distracter tasks that could realistically be expected to prevent covert rehearsal during these delay periods. One example in the free recall paradigm is Watkins, Neath & Sechler (1989). LIST PARSE predicts that serial recall with extended IOIs in which all rehearsals are prevented would preferentially decrease performance at the list beginning, thereby reducing primacy, but may improve performance at the list end. Increased IOIs cause the beginning items to experience competitive working memory interactions for much longer before recall begins, resulting in more compression of their activations and a loss of order information to noise. The extended IOIs between recent items result in their having more difference in their integration times, and thus more distinct activation levels and order retention, than they would with shorter IOIs. Thus, for short lists, longer delays between items can be beneficial to serial recall working memory performance, at least of the final few items, due to improved order encoding, in contrast to predictions of pure temporal decay accounts of working memory. Figure 17b summarizes simulations of the *order errors* predicted during serial recall of a six item list with ten simulation time unit IOIs, and no rehearsals. All other simulation procedures were identical to those used in the simulations of the Henson et al. (1996) data in Figure 7.

6. Conjunctive Coding of Order and Position

6.1. Cognitive Data. Various cognitive experiments suggest that position in an ordered sequence may be explicitly coded by the brain. For example, there is an increased likelihood of intrusions from items at similar list positions across trials (reviewed in Henson, 1998b). Repetitive presentation of an item in the same absolute serial position in lists across multiple trials with different list orderings leads to a statistically significant increase in the strength of learning of that item (Conway & Christiansen, 2001). There is a statistically small, but significant, class of errors called “protrusions”, in which interchanges between items in similar serial positions of different temporal groupings in a list are more common than interchanges between items in different serial positions (Henson, 1998b). That is, when presented with the sequence ABC (pause) DEF, interchanges between items at B and E are more common than interchanges between items at B and F. The well-known “slip of the tongue” error of “spoonerisms” illustrates a similar effect. Here, phonemes or syllables in similar positions in different words are selectively interchanged (e.g., “hissed my mystery lesson”). At present, models that do not assume explicit coding of position have not simulated these properties.

6.2. Neurophysiological Data. A growing number of neurophysiological experiments also report that individual neurons may be sensitive to absolute position of a sensory event or commanded motor action in a sequence. As noted in Section 3.14, cells have been identified that are selective for ordinal position. For example, a cell may respond to any target in the initial, middle or final ordinal position in a sequence, or the first, second, third, etc. positions. These cells are said to exhibit *rank order*, also called temporal selectivity or position specificity (Averbeck et al., 2003a; Barone & Jacobs, 1989; Funahashi et al., 1997; Inoue & Mikami, 2006: Kermadi & Joseph, 1995; Ninokura et al. 2004). Other cells show selectivity to *conjunctive* coding of item and ordinal position. Such a cell may respond to a specific target presented in a

specific ordinal position (Averbeck et al., 2003a; Funahashi et al., 1997; Inoue & Mikami, 2006; Ninokura et al., 2003, 2004; Shima & Tanji, 1998, 2000).

Given that Item and Order working memories can already explain many data about short-term storage of event sequences, what environmental problems supported the evolution of cells that also explicitly code positional information? What brain mechanisms enable item, order, and positional conjunctive coding, and could have arisen through evolution without requiring implausible types of specialized circuitry? The next section proposes answers to these questions, and outlines a framework wherein future versions of the LIST PARSE model can be consistently extended to include individual cells that conjunctively encode item, order, and positional information.

6.3. Position Coding, Repeated Events, Numerical Estimation, and Sequence Prediction. We have already seen that Item and Order working memories can arise from specialized recurrent shunting on-center off-surround networks that are ubiquitous in the brain (see Figure 2 and Appendix equation (2)). Cells in the simplest Item and Order working memories carry out conjunctive coding of item and order information, but the absolute position of an item in a list is not explicitly encoded. We propose that one source of positional information in the frontal cortex may be derived from the cortical maps of numerical representation that are found in the parietal cortex (Dehaene, 1997; Nieder and Miller, 2004; Pesenti et al., 2000; Pinel et al., 1999; Rickard et al., 2000), with projections to the frontal cortex (Nieder & Miller, 2003, 2004). We predict that such numerical maps contribute positional coding to the conjunctive coding of list item, order, and position that is found in higher cortical areas.

These numerical maps seem to have been involved in planning sequential behaviors for a very long time. They are, for example, important in foraging behaviors. They enable animals to make decisions about the total number of food items that are available in different locations, thereby guiding navigational and acquisitive behaviors (Buchanan & Bitterman, 1998; Emmerton, Lohmann, & Niemann, 1997; Washburn & Rumbaugh, 1991). Such maps support an abstract representation of number that can estimate the number of events in a sequence, even if they are presented via different modalities; e.g., seeing and hearing (Church & Meck, 1984; Starkey, Spelke, & Gelman, 1983).

How do such numerical representations arise? Grossberg and Repin (2003) have quantitatively modeled the properties of these numerical cortical maps. Their Spatial Number Network, or SpaN, model uses variations of more primitive cortical mechanisms for motion perception and spatial localization. This analysis provides an appealing answer to the question: How did numerical representation emerge during evolution? Where do numbers "come from"? These more primitive mechanisms are known to occur in the Where cortical stream, including the parietal cortex. In the brain's numerical maps and those of the SpaN model, distinct but overlapping cell populations are activated as an increasing number of events is detected. The model machinery that is sufficient to explain many numerical data properties is simple:

Successive events activate transient cell responses that are integrated through time (see Figure 18, *number of transient sensory signals*, and *integrator amplitude*). This integrated activity grows with the number of events that are being counted. The integrated activity is uniformly broadcast (see Figure 18, *uniform input*) to a network of cells that interact with one another via a recurrent shunting on-center off surround network. The cells at one end of the network have signal functions that respond to these uniform inputs with lower thresholds and less sensitivity; that is, their signal functions have flatter slopes as a function of cell activity. The cells at the other end of the network have higher thresholds but greater sensitivity; that is, their

signal functions have steeper slopes as a function of cell activity. Cells in between have intermediate thresholds and sensitivity (see Figure 18, *transfer functions with variable thresholds and slopes*).

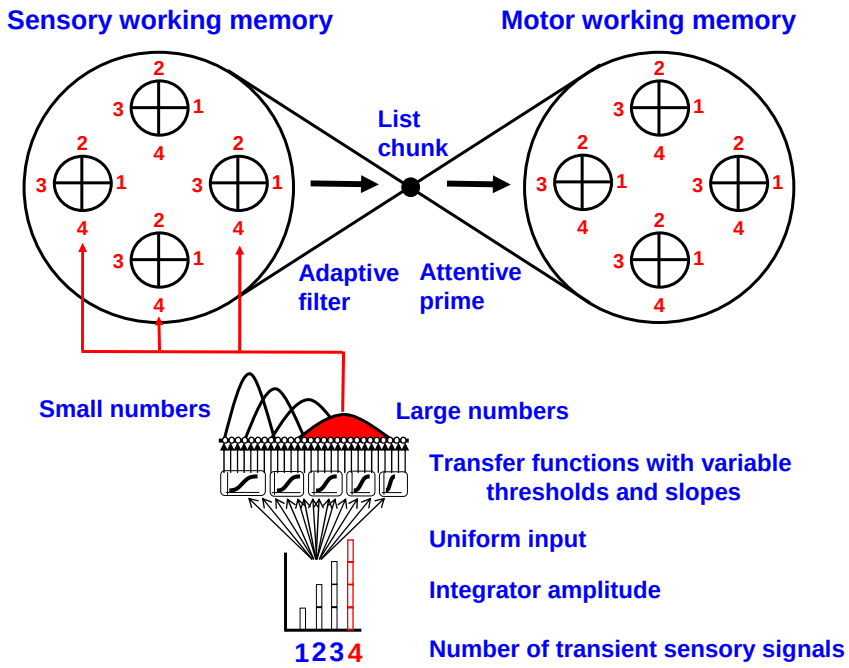


Figure 18. Circuit for encoding a conjunction of item, order, and position in sensory and motor working memory. Cells in the sensory and motor working memories need a second input that codes positional information in order to fire. The model proposes that number maps of the type found in parietal and frontal cortex provide this positional information. The circles with numbers represent cortical hypercolumns, each coding a different sensory or motor event, with positions (for illustration) 1, 2, 3, and 4. The sensory working memory is unitized through learning by one or more sequence-sensitive list chunks. The list chunks learn to attentively prime the item-order-position motor working memory during reactive performance of a sequence of actions, while the list chunk is active. During planned performance, cells in the motor working memory fire their motor commands when they receive a list chunk priming signal and the correct positional input from the corresponding number map. The lower part of the figure illustrates how transient inputs that register when a sensory event occurs are integrated into a signal proportional to the total number of sensory inputs that have occurred in the sequence. This integrated signal generates a uniform, growing input to all the cells in the the number map. The signal functions with variable thresholds and slopes in the number map cause distinct populations of cells to get activated as a larger number of transients is stored. These number map cells broadcast their positional information to the sensory working memory. A similar scheme occurs in the motor working memory. See text for details.

This gradient of threshold and sensitivity creates a Position-Threshold-Slope (PTS) Shift Map (Grossberg & Kuperstein, 1986/1989). In any such map, all the cells receive the same input, which increases with the total number of inputs in the sequence. The cells then compete via their

signal functions with different thresholds and slopes. After the competition takes hold, the position of maximal sensitivity shifts to the right as the total input increases (see Figure 18, *small numbers* and *large numbers*). This happens because, for small inputs, cells at the left of the map get activated due to their smaller thresholds. As larger inputs arrive, cells increasingly to the right of the map also get activated when their larger thresholds are exceeded. Since these latter cells have greater sensitivity (slopes) to input increments, their output signals grow faster, so that they can inhibit the cells corresponding to the smaller inputs more effectively than conversely. Hence, the peak of maximal activation shifts to the right. In the present case, this PTS Shift Map is a number map because cells respond selectively to larger numbers as the uniform input that all cells receive grows with the total number of events.

We propose that cortical cells that conjunctively code item, order, and positional information in a list may be the result of (at least) two interacting neural designs: specialized recurrent shunting on-center off-surround networks that form an Item and Order working memory, and specialized recurrent shunting on-center off-surround networks that support an ordered number line, and can thus represent the absolute position of an event in a list (Figure 18).

Assuming, as will be shown below, that interactions of these two types of circuits lead to cortical cells that can conjunctively code item, order, and list position, one still needs to ask what additional properties can be realized by these circuits that are harder to realize using only the simplest Item and Order working memories? One such property is the ability to efficiently code item *repetitions* in a sequence. Another such property concerns the ability for a single list chunk to *predictively prime performance of an entire sequence* of future sensory expectations or actions, rather than just the next expectation or action. Before describing how the circuit in Figure 18 can conjunctively code item, order, and list position, we briefly review related earlier models that contain concepts and mechanisms that can be refined to realize this desired result.

An earlier Item and Order working memory model, called the STORE model (Bradski, Carpenter, and Grossberg, 1994). has shown how lists with repeated items can be stored in working memory. In this model, each item in a list can activate a cortical *hypercolumn*, or mini-map, in which different cells in the hypercolumn can represent distinct occurrences of the same item when it is repeated in a list. The total Item and Order working memory activity gradient can then represent the entire list, including item repetitions, in their correct order, up to the maximal length of the stored primacy gradient. However, this sort of hypercolumn map does not represent the absolute position of an item in the list.

The model in Figure 18 also uses hypercolumns that can represent repeats of a given sensory cue or action in a list. By using inputs from the cortical number map, this type of circuit can also code absolute item position within a list. For example, consider the *sensory working memory* in Figure 18. When, say, item 4 in the list is presented, this number input is broadcast to all the hypercolumns in the sensory working memory; that is, the projections labeled 4 in Figure 18 all get activated when the fourth list position is active. Such number map activation may occur just as it does to count events that occur across modalities (Church & Meck, 1984; Starkey, Spelke, & Gelman, 1983). The only hypercolumn cell that can get activated corresponds to the sensory cue that occurs in position 4 of the list. Occurrence of such a sensory cue *primes* all the cells in its hypercolumn. Both cue and temporal order inputs need to converge on a cell in the sensory working memory for it to become fully active.

The concept of predictive priming has a long history in neural modeling. The *context-sensitive avalanche* model (Grossberg, 1974, 1978a) is particularly relevant to conjunctive coding. In such a circuit, a context-sensitive chunk can prime the activation of a list of planned

expectations or actions. Although all cells that code the plan are primed, the execution of a preceding command is needed before the next event can be fully activated. This circuit does not code the absolute position in a list, and does not easily code repeated performances of the same item at multiple positions of a list.

By using a SpaN number-based map as the inputs to both sensory and motor working memories, both the repetition and prediction problems can be solved using the same neural mechanisms, while also coding the absolute position of each item in a list (see Figure 18). Here is how that can be achieved:

Hypercolumns and Repeats. In the STORE model of how repeated events can be stored in working memory, Bradski, Carpenter, & Grossberg (1994) proposed that each item in a list can activate a cell population, or hypercolumn, across which a gradient of asymmetric competition occurs. These cell populations are part of a larger recurrent shunting on-center off-surround network that comprises a working memory. The first occurrence of the item activates some cells in this population more than others, because of the asymmetric competition. Later occurrences of the item are able to activate a distinct set of cells in the population. And so on, until the whole population is storing multiple repetitions of the item in the list. Each hypercolumn thus conjunctively codes item identity and the relative positions of that item's occurrence in the list. The total item and order working memory gradient, across all of the active hypercolumns, can then represent all the items and their repetitions, in their correct order, up to the maximal length of the stored primacy gradient. This kind of conjunctive coding can store item repeats in their relative order, but it does not encode the absolute position of each item in the list.

Conjunctive Coding of Item, Order, and Position. The model in Figure 18 can represent item, order, and absolute list position by combining hypercolumns with inputs from a cortical number map. As above, different cells in each hypercolumn can store repeats of a given sensory cue (Figure 18, *sensory working memory*) or action (Figure 18, *motor working memory*). A number map enables absolute list position to be encoded when it inputs redundantly to multiple hypercolumns that correspond to individual sensory items or motor actions (Figure 18). The main idea is that each sensory or motor representation is represented by a cell population, or hypercolumn. Different cells in each hypercolumn are activated by different numbers, or absolute list positions, in the number map.

For example, consider the sensory working memory in Figure 18. If full activation of a cell in sensory working memory requires its sensory input and an input from the number map, then the hypercolumn as a whole can represent a sensory event with multiple repeats, each of which is activated by the number corresponding to the item's absolute position in the list. When, say, item 4 in the list is presented, this number input is broadcast to all the hypercolumns in the sensory working memory (see projections labeled 4 in Figure 18). The only cells that can get activated correspond to the sensory cue that occurs in position 4 of the list.

The sensory-working memory learns to activate a list chunk, or chunks (Figure 18), just as it does in Figures 2 and 11. Such a chunk encodes the item, order, and positional information of the entire sensory list. When activated, these sensory list chunks can prime all the cells in the motor working memory that can command actions in the correct order (Figure 18). When the correct number, or list position, input is received by the motor working memory, the corresponding motor command is released.

Are multiplexed hypercolumns known to occur in the brain? The answer is well-known to be "yes". They are particularly well studied in visual cortex, including hypercolumns that code position, orientation, and ocular dominance in cortical area V1 of visual cortex (Blasdel, 1992a,

1992b; Hubel & Wiesel, 1968); and hypercolumns that code position and direction of motion in cortical area MT (Albright et al., 1984; Maunsell and van Essen, 1983). It is also known that combinations of sensory events and reinforcers are represented in even higher cortical areas, notably orbitofrontal cortex (Rolls, 2000, 2004; Schoenbaum & Roesch, 2005; Schoenbaum & Setlow, 2001) and may enable selective decisions to be learned in response to the same sensory cue under different motivational and reward conditions. Such a conjunction is predicted to occur in the orbitofrontal cortex of the CogEM model (see Section 2.2 and Figure 3).

The proposed conjunctive item and position sensitivity in working memory is assumed to be yet another example of this ubiquitous brain design (Averbeck et al., 2003a; Funahashi et al., 1997; Inoue & Mikami, 2006; Ninokura et al., 2003, 2004; Shima & Tanji, 1998, 2000). The activation gradients across these conjunctive representations are assumed to arise from the usual recurrent shunting on-center off-surround network dynamics, just as they do in other working memory networks.

Hypercolumn representations that include number-order constraints may exist in prefrontal cortex for the same reasons that they exist in visual and orbitofrontal cortex (see Section 6.2); namely, to fit multiple constraints (e.g., position, orientation, ocular dominance) within a two-dimensional cortical surface. However, in cases where the sensory or motor representations are one-dimensional (e.g., movement directions), then it is also possible that the feature and number lines may realize their conjunctive codes using a simpler cross-bar circuit wherein parallel cortical strips represent each feature, and parallel number strips represent each number, with the two sets of strips intersecting in cells that require a pair of strips to be active to fire the conjunctive cells.

A cross-bar circuit for conjunctive coding has been predicted by Grossberg and Repin (2003) to explain *place value* number codes—e.g., hundreds, thousands, millions, etc.—which provide the human number system with its open-ended properties. These codes are predicted to arise through learned associations from categorical language representations in the temporal and frontal cortex to the spatial representation of the number line in the parietal cortex. An auditory language category (e.g., hundred) primes a learned strip of parietal cortex which represents all the numbers in the number line. Inputting a specific number (e.g., two) to the number line activates the cortical strip that corresponds to that number. Where these two strips intersect, a conjunctive map representation of a specific number (e.g., two hundred) is activated.

Context-sensitive avalanches. A complete model of positional coding needs to incorporate other mechanisms of the *context-sensitive avalanche* architecture, which can now be identified. Such a model is a natural development of avalanche concepts through the years.

An early neural model of serial order in behavior analyzed serial verbal learning data in terms of inter-item learned associations and competition (Grossberg, 1969, 1978a; Grossberg & Pepe, 1971). This model explained and simulated various classical serial verbal learning properties, including the bowed serial position curve and the asymmetric distribution of anticipatory and perseverative errors at a list beginning, middle, and end. The model also led to the realization that the unitized representations that are linked by such associations should be list chunks, not individual items, and that such list chunks arise through learning as categories that are selectively activated via a bottom-up adaptive filter by sequences of events as they are stored temporarily in working memory. This realization led both to the introduction of Item and Order working memories, and to the introduction of Adaptive Resonance Theory, to explain how lists can be temporarily stored and chunks can be learned in a stable way (Grossberg, 1978a, 1978b).

This combination of insights led to the introduction of the context-sensitive avalanche model in Grossberg (1974, 1978a). This model includes working memory gradients, the list chunks that they activate, inter-chunk associations, and the output commands of the list chunks. The "avalanche" consists of the associative links among the list chunks. The context-sensitivity derives from the fact that the associative links are learned between context-sensitive list chunks, rather than low-level features. In addition, higher-level planning chunks may learn to prime all the list chunks of an avalanche before volition releases overt performance of the sequence.

In such an avalanche, as well as in LIST PARSE, list chunks are learned in response to working memory gradients, and can read-out both top-down expectations to reactivate the gradients, and output motor commands. LIST PARSE is not a complete context-sensitive avalanche because it does not include the inter-chunk associations. However, the model machinery is available to be incorporated into a future, more comprehensive, version of LIST PARSE. In particular, list chunks in LIST PARSE occur in cortical layer 2/3. Inter-chunk associations are predicted to be homologs of the long-range horizontal connections that are found in layer 2/3 of a visual cortical area like V2, where they have been modeled to carry out long-range perceptual grouping; cf., Grossberg (2003a).

Conjunctive coding is a property of even the earliest avalanche models (Grossberg, 1974). In these models, the list chunk is a command cell that primes activation of all the cells that will control sequential performance. Such command cells exist even in invertebrates. As noted above, order information is represented using inter-item associations that, together with chunk activation, can activate the next command after the preceding item or action has been activated. In Figure 18, the list chunk plays the role of the command cell, and the number-sensitive map allows the next event to be triggered without the need for learned inter-item associations. To complete the design of context-sensitive-avalanches-with-position-coding, learned associative links occur between the list chunks to help select the next chunks to be activated through time. A full development of how context-sensitive avalanches govern cognitive learning, memory, and planning in both normal and clinical human populations is a future goal of this research.

6.4. Reset Events and Positional Codes. In order for position in the list to be coded, the number representation that activates list position must be updated, or reset, as each sensory event occurs, or as the previous motor action is almost finished. If such a reset event generates a transient burst of activity to the number map integrator, then the number map increases its count by one, and the next sensory or motor item can be activated by satisfying the conjunctive constraint.

Such reset events are familiar in the neural modeling literature. For example, one index of processing a visual cue, or of completing an eye movement, is foveation of the cue. Fixation cells can be activated in multiple brain regions, such as superior colliculus and frontal eye fields (e.g., Munoz & Wurtz, 1995), when the eyes fixate a visual cue. They help to maintain foveation of the cue while it is stationary. This raises the question of whether activation of fixation cells can also cause a transient input to the integrator that feeds a number map. Arm movement controllers also compute indices capable of determining when a movement is almost complete, such as the mechanism, described in Section 4.4, that controls anticipatory selection of the next movement in a sequence.

6.5. Spoonerisms and Other Positional Errors. Although Figure 18 shows only one list chunk for simplicity, multiple list chunks can simultaneously be active, as in a masking field (see Section 2.4), each with a different activation level corresponding to its predictive relevance in a given spatio-temporal context. When two or more chunks are active, and share a number of

overlapping features, then the same number, or list position, can activate more than one list's output in that position. Moreover, when that output is triggered, self-inhibition of its item representation can prevent its performance while a second sequence is quickly performed. This combination of multiple list-chunks, number line output triggers, and inhibition of return opens the way to natural explanations of many properties like spoonerisms and related properties that are summarized in Section 6.1.

7. Comparative Analysis of Working Memory Models.

7.1 Free Recall Models: Distinctiveness and Dual-Store Models. Free recall models need to explain how subjects use order information during performance, and free recall rehearsal strategies. The simplest strategies, such as repetitively rehearsing in serial order from the start of the list, or repetitively rehearsing the last presented items, are not necessarily either optimal or the way that subjects actually rehearse (e.g., Rundus, 1971; Tan & Ward, 2000).

Temporal distinctiveness models (e.g., Bjork & Whitten, 1974; Brodie & Murdock, 1977; Brodie & Prytulak, 1975; Crowder, 1976; Howard & Kahana, 2002; Murdock, 1960; Nairne, 1990; Rundus, 1971; Tan & Ward, 2000) typically propose that the ratio of the time between instances of item rehearsals (e.g., between the first and last rehearsal, ΔT; see Figure 9f) and the time from the last rehearsal of a particular item to the beginning of recall, T (Figure 9d), largely accounts for the item's likelihood of recall. An analogy that illustrates the distinctiveness model (e.g., see Tan & Ward, 2000) is a line of telephone poles with inter-pole distance, ΔT, with distance T between the observer and the last pole. Choosing ΔT smaller or T larger makes it more difficult to distinguish any particular pole. This is an appealing metaphorical explanation of free recall data. However, it provides no mechanistic interpretation of why a distance metaphor may be appropriate, and has not been generalized to the ISR task.

LIST PARSE working memory mechanisms provides the first neural explanation of the temporal distinctiveness "ratio rule" of Tan & Ward (2000) of which we are aware: For all but the last rehearsed, or presented, items in a span or super-span list, the longer an item goes unrehearsed in the presence of other items, the less likely it is to be recalled, because competitive normalizing dynamics continually push items toward a sub-threshold equilibrium point. This tendency also correlates with the detrimental effect of increasing "retention interval (T)" in the ratio rule of Tan & Ward (2000). With items rehearsed at the ends of long lists, however, retention times are short before their recall. Strong recency effects occur because these recently presented items have had less time to decay, and still retain much of the activation they acquired at last presentation; cf., the recency gradients in Figures 9a and 12c.

Rehearsing an item in working memory boosts its activity relative to the other items in the list. This boosted activity makes it more likely that this item will be rehearsed again later in the list. Items at the beginning of lists are therefore rehearsed more often and further down the list, relative to their starting position, than other items (see Figure 9f), because fewer items compete with them to be rehearsed early in the list, and these early rehearsals boost their representations relative to the items that come in the middle of the list. Therefore, extended primacy effects can be established due to extensive rehearsal of the first items in the list. These effects roughly correlate with the "ΔT" effects of the ratio rule of Tan & Ward (2000).

Rehearsal grouping can influence probability of recall (Tan & Ward, 2000) and has a clear explanation in LIST PARSE: Due to the competitive dynamics of the cognitive working memory, the relative boost in activation due to the rehearsal of an item that is already highly active is less than that for an item with lower activation. Therefore, several rehearsals of items that are temporally close together will have less net effect in keeping their activity large, and thus

in increasing their probability of eventual recall, than the same number of rehearsals more evenly spaced throughout the list.

Dual-store models (e.g., Davelaar, Haarmann, Goshen-Gottstein, Ashkenazi & Usher, 2005; Davelaar, Haarmann, Goshen-Gottstein & Usher, 2006) also have neural homologs and extensions in LIST PARSE. Devalaar et al. proposed that free recall performance is mediated by two factors: a recency-dominated short-term memory of recurrent on-center off-surround leaky competitive nodes, whose dynamics are similar to the antecedents of LIST PARSE in Grossberg (1973, 1978a, 1980, 1988), and learned long-term memory connections between these nodes and a context representation.

Comparison of LIST PARSE to classical cognitive models of free recall probability, like SAM (Raaijmakers & Shiffrin, 1980), ACT-R (Anderson, Matessa & Lebiere, 1997; Anderson, Bothell, Bebiere & Matessa, 1998), or the model of Howard & Kahana (1999, 2002), is difficult because LIST PARSE currently does not currently include the inter-item associations, or recall from episodic context cues, that are used by these models. In addition, these models do not propose testable neural mechanisms.

Because of this omission, several types of free recall data have yet to be quantitatively simulated by LIST PARSE. For example, the *lag-recency effect* (Kahana & Howard, 2005; Klein et al., 2005) describes the conditional probability of the *next* item to be recalled during the recall period of the free recall paradigm. Inter-item associative and/or episodic contextual cuing may be involved in this ordering. A full analysis of such data would include both the learned list chunks that LIST PARSE does simulate, and inter-chunk associations that it does not. The context-sensitive avalanche circuits that are summarized in Section 6 will help to explain many such data in future modeling studies.

7.2 Serial Recall Models. Neural modeling approaches to serial recall, particularly ISR, can generally be placed into three schools: associative chaining, positional, and ordinal (Henson, 1998b).

7.2.1. Associative Chaining Theories. Associative chaining theories, for example TODAM (Lewandowsky & Murdock, 1989), the Jordan (1986) compound-chaining model, or the Elman (1990) simple recurrent network (SRN; see also Botvinick & Plaut, 2006a, 2006b) model, rely principally upon associative links between ordered items to produce serial responses. While inter-item associations may play a role in list learning across multiple trials (e.g., Klein et al., 2005), and these models are conceptually simple, their mechanisms are inconsistent with several types of data that involve working memory and sequential performance. Indeed, they are faced with all the problems that led to the introduction of Item and Order working memories as a possible way out.

Perhaps the most famous counter-evidence for an associative chaining account of serial performance is the Lashley (1951) analysis of typing errors and kinematics, in which he found that his own typing errors tended to be transpositions of neighboring items (for a more rigorous analysis of typing errors, see MacNeilage, 1964). By analyzing the response times of experienced typists, he concluded that feedback signals from a completed key press could not signal the next key press. Lashley concluded that nerve fiber signal conduction is too slow and many actions must have central parallel "partially activated or readied" representations that can be anticipatorily queued. For a more extensive review of limitations of associative chaining and possible alternatives, see Hartley & Houghton (1996) and Henson et al. (1996).

In response to these well-accepted limitations, all modern chaining theories incorporate central planning mechanisms. As noted in Section 6.1, Item and Order working memories were

themselves derived from an analysis of limitations of associative chaining theories and of how working memory gradients, list chunks, and inter-chunk associations could be joined together for cognitive and sensory-motor control into context-sensitive avalanches (Grossberg, 1978a).

Another set of data that challenge purely associational chaining accounts of serial recall is *filling-in* (see Section 3.6). Chaining models assume asymmetric weight strengths with stronger associations from an item to subsequent items than to preceding items. Associative asymmetry can be mathematically proved to emerge naturally from associative learning laws when they are combined with competitive dynamics, and has been applied to the explanation of serial verbal learning data (Grossberg, 1969, 1978a; Grossberg & Pepe, 1971). However, such a mechanism cannot be the whole story. For example, when an error is made where an item is skipped, asymmetric associative strengths should make it more likely for the system to continue in a forward direction during list recall, rather than moving backward and picking up the skipped item, as is observed. It is here that working memory gradients play a big role.

Associative chaining models also make inaccurate predictions with respect to repetition phenomena; for instance, the *Ranschburg effect* (Henson, 1998a; Howard & Kahana, 2000): When repeated items in a list are separated by more than one item, a repeated item is less likely to be recalled than a distinct item. If there is less separation, then the repetition is more likely to be recalled than a distinct item. Also, there is no observed impairment in recalling the items after the repeated items. Chaining models would predict that selection of a repeated item should cue both of the distinct items that it predicts, but this is not observed.

7.2.2. Positional Theories. Positional theories (e.g., Burgess, 1995; Burgess & Hitch, 1992, 1999; Hartley & Houghton, 1996; Henson, 1998b; Johnson, 1991) posit that positional information is explicitly maintained to establish recall order. Such positional information may be absolute, or relative to the beginning and/or end of lists.

Positional models help to explain data wherein there is an increased likelihood of intrusions from items at similar list positions across trials (reviewed in Henson, 1998b), and repetitive presentation of an item in the same absolute serial position in lists across multiple trials with different list orderings leads to a statistically significant increase in the strength of learning of that item (Conway & Christiansen, 2001). Also, cognitive data (Henson, 1998b) have indicated the presence of a statistically small, but significant, class of errors called “protrusions”, in which interchanges between items in similar serial positions of different temporal groupings in a list are more common than interchanges between items in different serial positions. That is, when presented with the sequence ABC (pause) DEF, interchanges between items at B and E are more common than interchanges between items at B and F. The well-known “slip of the tongue” error of “spoonerisms” illustrates a similar effect. Here, phonemes or syllables in similar positions in different words are selectively interchanged (e.g., “hissed my mystery lesson”). At present, models that do not assume explicit coding of position have not simulated these properties.

Common weaknesses of these models include: The positional information in the models is often difficult to biologically implement or justify based upon observed physiological mechanisms. Moreover, the order information often does not emerge from a self-organizing process. For example, some models require knowledge that an item is the end of a list in order to accurately calculate the end-of-list positional cues (Henson, 1998b), while others use attentional windows which presumably must learn to accurately apply relative positional information with the window across a wide, and unpredictable, spatial distribution of cells (Burgess & Hitch, 1999). Available models do not scale well to account for multi-level serial order storage. For example, they require the use of attentional windows to provide positional information (Burgess

& Hitch, 1999), which implies that there must be analogous low-level attentional windows to provide the same information and allow the learning of sequential information at lower-levels in the hierarchical learning process. Finally, they do not adequately treat item repetitions. For example, they do not account for both the error and inter-response time properties of the Ranschburg effect (see Kahana & Howard, 2000, for a review of these limitations).

Botvinick & Plaut (2006a, 2006b) have proposed a recurrent network model of immediate serial recall that simulates a number of data from this paradigm. In the model, "item and position are coded conjunctively" (p. 207) in such a way that "some hidden units code for item (independent of position) others code for position independent of item), and still other code exclusively for a single item-position conjunction. However, further analysis indicated that the vast majority of hidden units showed coding properties intermediate among these extremes" (p. 232). In fact, "very few, if any, units coded in an exclusive manner for a single item–position conjunction" (p. 233).

The authors cited their use of complex distributed representations of items, where units partially code for multiple features, as a fundamental difference with models using localist representations, where individual model units are associated with the item content or responses for which they code (cf. the Primacy model of Page & Norris (1998) and LIST PARSE). Page (2000) extensively addressed weaknesses of the widespread model use of "distributed" representations, and we specifically address a few of these points below. However, almost all models, including LIST PARSE and Botvinck & Plaut (2006a), use a mixture of localist and distributed representations at different processing stages. Thus, general labels like "localist" or "distributed" often represent a distinction without an essential difference.

Much like other positional models to date, the model hypotheses that are used to generate positional sensitivity in Botvinct & Plaut (2006a) are not supported by behavioral or neurobiological correlates. One concern is that the model used very slow learning to create the network representation that was used to fit the data. The hypothesis of slow learning is contradicted by the very notion of *immediate* serial recall, in which subjects are required to recall items after a single exposure to a novel list. Another concern is that the learning in question is supervised by "teacher forcing" that drives the network structure to create desired outcomes. In particular, "the activations propagated over the recurrent connections from output to hidden layers were based on the target values for the output units, not their actual activation" (p. 206). Teacher forcing is a steepest-descent curve-fitting procedure whose success says little more than the model had enough degrees of freedom to fit this desired outcome.

A further concern about the supervised learning protocol is that "training proceeded until the network reached a predetermined level of recall accuracy (proportion of lists of a selected length recalled correctly). For this purpose, performance was evaluated after every 10,000 trials. The reference accuracy level was drawn from the relevant empirical studies…" (p. 206). These externally controlled manipulations have no analog immediate serial recall tasks. The model's results are, moreover, dependent upon the exact training schedule: "training the model only on a single list length was found to produce results qualitatively different from those that are reported here…list length increased by one element per trial until a simulation-specific maximum length was reached. Following presentation of a list of maximum length, the list length returned to one and the cycle repeated" (pp. 205-206). Finally, because the learning mechanism is back propagation, it uses non-local learning operations that have no analog in brain dynamics (cf., Grossberg, 1988). Because model representations that are derived by non-local, slow, teacher-forced supervised learning cannot be compared with the way in which immediate serial recall

experiments are carried out, and have no known analog in brain mechanisms, it is hard to justify using these learned model representations as a foundation for explaining the order and positional representations found in the brain.

The central advantage cited by Botvinick & Plaut (2006a) for their learned representations was to explain similarity effects in serial recall using a single hidden layer stage. This stage is used to explain why, when a serial list includes a subset of items that are easily confusable, these items have higher error rates due to transpositions between confusable items, thereby creating a "saw tooth" error pattern. In contrast, the Primacy model of Page & Norris (1998) assumes a second stage of processing after serial order selection at which item similarity confusions occur, leading to the observed selective transpositions. Although these similarity effects have not been quantitatively simulated herein, the LIST PARSE model is compatible with a two-stage process in which one cause of confusability occurs during phonemic reconstruction of verbal output (see Section 7.2.3). At present, it is unclear how many stages are involved in the brain. It is also unclear that a stage of hidden units in a back propagation model with convolved distributed representations offers a parsimonious cognitive explanation, particularly given that a regression analysis is needed to determine the factors that contribute to a given unit's activity (Botvinick & Plaut, 2006a, p. 232).

Some recent positional models (e.g. Henson, 1998) use item position cue gradients to maintain positional information: Primacy gradients represent position relative to the list start, and recency gradients positions relative to the list end. One may question whether such relative positional models should be classified as positional versus ordinal models.

Two distinctions may offer a more meaningful classification: "absolute versus relative order coding" and "intrinsic versus explicit order coding". If the value of the model property responsible for coding order, in isolation from other items' values, can be assessed to determine its order in a list, then it uses "absolute" order coding. If comparison with the value of other items is necessary to determine its order, then the model uses "relative" order coding. For instance, models that propose positional slots for the 1st, 2nd, …, n^{th} items in a list (Atkinson & Shiffrin, 1968) are absolute. The current versions of LIST PARSE and Start-End (Henson, 1998b) models are relative models because only by comparing the activity (or weight) levels of two items can one determine which will be recalled later in the list. If the mechanism that determines its ordering (e.g., activation level) is proposed to serve functions other than temporal order storage, it is an "intrinsic" property, whereas otherwise it serves as an "explicit" order cue. Item and Order working memories use an "intrinsic code of order" by proposing that item activity codes both item and order information. The Start-End model gradients are exclusively codes of item order and not proposed to have any other model role, and are thus an "explicit code for order".

7.2.3. Ordinal Theories. LIST PARSE, in its simulated form herein, belongs to a third class of models, ordinal theories (e.g., Bradski et al., 1994; Cohen & Grossberg, 1987; Farrell & Lewandowsky, 2002; Grossberg, 1978a, 1978b; Nigrin, 1993; Page, 1994; Page & Norris, 1998), which represent order on the basis of a *relative*, continuous, *intrinsic* property of the items themselves, rather than as *explicit* positional cues. Ordinal models have several important properties, including: They account for certain behavioral data sets (e.g., response latency data; see Farrell & Lewandowsky (2004) for a review) that other models cannot explain at present. They use conceptually simple, self-organizing mechanisms that are ubiquitous in the brain (e.g. neural firing rates, and recurrent on-center off-surround interactions). Finally, they have received direct support from neurophyiological data (e.g., Averbeck et al (2002, 2003a, 2003b).

Ordinal models are also currently incomplete: They have not yet simulated all the important item repetition phenomena; e.g., the Ranschburg effect. However, Bradski et al. (1994) describe an Item and Order working memory that is capable of storing repeated items, and in which items separated by more than one item are less likely to be recalled. The same is true for the Item and Order working memory with positional coding that was summarized in Section 6. The Henson (1998a) proposal that repetitions separated by 0 or 1 items are stored as a single chunk is compatible with such a model.

LIST PARSE, even in its current form without explicit simulations of positional information, qualitatively explains and neurally interprets a much larger set of data than are quantitatively simulated in this first article about it. Some of these data have been simulated by other models that use Item and Order working memory properties. Responses latencies in serial recall (e.g., Farrell & Lewandowsky, 2002, 2004; Sternberg et al., 1978) are clearly compatible with LIST PARSE, as related Item and Order models have already shown; e.g. Rhodes et al. (2004). Several types of serial recall cognitive data were simulated by the primacy model of Page & Norris (1998; also see Henson, 1998b), which assumes a primacy gradient that LIST PARSE generates as an emergent property. In the primacy model, the effects of phonemic similarity of items were simulated using a second processing stage wherein phonological codes are reconstructed for motor output.

A mechanistically more articulated description of such a process occurs in LIST PARSE, where selected speech plans S_i are transformed into speech motor targets T_i in the trajectory generator (Figures 2 and 14). These speech motor targets comprise a group of motor features, reflected biologically as a cell population, which partially overlap. This overlap reflects the degree of item similarity, such that different planned words with shared speech sounds can each activate cells representing those shared sounds. As a result, selection of a plan primes motor targets that are associated with similar plans. "Saw tooth" error patterns naturally follow from this stage, in which phonologically confusable items are less likely to be recalled in their correct serial position due primarily to transpositions with other phonemically similar items. Chunking properties (Bower, 1972), such as the ability to recall longer lists of letters if they are familiar (e.g., FBI or CIA), naturally follow from properties of list chunks in masking fields (Grossberg, 1978a). Finally, components of the LIST PARSE macro-circuit in Figure 1 that are not underlined have elsewhere been modeled, as reviewed in Section 2. These components will be incorporated to explain larger data sets in future modeling studies.

8. Concluding Remarks. The LIST PARSE model embodies both qualitative and quantitative advances beyond previous ordinal and competitive queuing models, including the following properties:

LIST PARSE shows how working memory gradients emerge and change dynamically through time, including bowed gradients with extended recency portions, and how such gradients can support on-line learning and read-out of list chunks. The interactions between these short-term and long-term memory processes enables LIST PARSE to begin to unify several different types of data that have heretofore not yet received a neurally plausible mechanistic explanation. In particular, LIST PARSE also offers the first unified quantitative neural explanation of human cognitive data about both immediate serial recall and immediate free recall. It proposes how and why similar working memory dynamics can generate these data when combined with different rehearsal strategies.

Along the way, LIST PARSE predicts neurobiologically testable mechanisms for how distinct cognitive and motor working memory circuits work together in ventrolateral and

dorsolateral prefrontal cortex; proposes how volitional modulation of cognitive and motor working memory circuits by the basal ganglia can control variable-rate, anticipatory, sequential performance; is the only model to quantitatively simulate both human cognitive data and monkey neurophysiological data about working memory storage and sequential performance; and learns list chunks that are capable of learning to categorize and store in long-term memory event sequences of variable length. The masking fields which model variable-length list chunks have earlier been used to quantitatively simulate challenging human cognitive data about contextual effects during speech perception and word recognition (Grossberg & Myers, 2000). These earlier studies did not, however, simulate how the list chunks are learned. LIST PARSE provides a foundation for learning such list chunks through language experiences and then simulating speech perception and word recognition data as consequences of this language learning process;

LIST PARSE may also be naturally extended to provide conjunctive codes of item, order, and position in a list. Together with number-based cortical maps to provide the positional sensitivity (Grossberg & Repin, 2003), and context-sensitive avalanche inter-chunk associations (Grossberg, 1974), this extended architecture promises to explain all the types of data that are summarized in this article.

Last but not least, LIST PARSE advances an emerging unified theory of how laminar neocortical circuits support all higher intelligent behaviors. It shows how variations of the LAMINART circuits that have previously helped to explain many data about visual perception and neuroscience (e.g., Grossberg, 2003a; Raizada and Grossberg, 2003) can also explain data about human and monkey working memory, list learning, and sequential performance. Taken together, the LAMINART and LIST PARSE models make predictions about how granular neocortical circuits may be specialized to learn and carry out both spatial and temporal behaviors, including how variations of the same volitional mechanisms can control visual imagery and fantasy as well as storage of event sequences in cognitive working memory.

REFERENCES

Agam, Y., Bullock, D., & Sekuler, R. (2005). Imitating unfamiliar sequences of connected linear motions. *Journal of Neurophysiology*, *94*, 2832-2843.

Agam, Y., Galperin, H., Gold, B. J., & Sekuler, R. (2006). Learning to imitate novel motion sequences. Under review.

Albright, T. D., Desimone, R., & Gross, C. G. (1984). Columnar organization of directionally sensitive cells in visual area MT of the macaque. *Journal of Neurophysiology, 51*, 16-31.

Aldridge, J. W., & Farrell, M. T. (1977). Long-term recency effects in free recall. *American Journal of Psychology, 90*, 475-479.

Alexander, G. E., & Crutcher, M. D. (1990a). Preparation for movement: neural representations of intended direction in three motor areas of the monkey. *Journal of Neurophysiology*, *64*, 1, 133-150.

Alexander, G. E., & Crutcher, M. D. (1990b). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences*, *13*, 7, 266-271.

Alexander, G. E., Crutcher, M. D., & DeLong, M. R. (1990). Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Progress in Brain Research*, *85*, 119-146.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357-381.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*, 341-380.

Anderson, J., Matessa, M., & Lebiere, C. (1997). ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention. *Human-Computer Interaction*, *12*, 439-462.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 89-195). New York: Academic Press.

Atkinson, R.C., & Shiffrin, R. M. (1971). The control of short term memory. *Scientific American*, *225*(2), 82-90.

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences, USA, 99*, 20, 13172–13177.

Averbeck, B. B., Crowe, D. A., Chafee, M. V., & Georgopoulos, A. P. (2003a). Neural activity in prefrontal cortex during copying geometrical shapes. I. Single cells encode shape, sequence, and metric parameters. *Experimental Brain Research, 150*, 127-141.

Averbeck, B. B., Crowe, D. A., Chafee, M. V., & Georgopoulos, A. P. (2003b). Neural activity in prefrontal cortex during copying geometrical shapes. II. Decoding shape segments from neural ensembles. *Experimental Brain Research*, *150*, 142–153.

Averbeck, B. B., & Lee, D. (2007). Prefrontal neural correlates of memory for sequences. *Journal of Neuroscience*, *27*, 9, 2204-11.

Averbeck, B. B., Sohn, J., & Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neuroscience*, *9*, 2, 276-282.

Avons, S. E., Wright, K. L., & Pammer, K. (1994). The word-length effect in probed and serial recall. *Quarterly Journal of Experimental Psychology*, *47A*, 207–231.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *Recent Advances in Learning and Motivation (Vol. 8)*, New York: Academic Press.

Baddeley, A. D., & Scott, D. (1971). Short-term forgetting in the absence of proactive interference. *Quarterly Journal of Experimental Psychology*, *23*, 275-283.

Baddeley, A. D., Thompson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575-589.

Baddeley, A. (1986). *Working Memory.* London: Oxford University Press.

Barbas, H. (1992). Architecture and cortical connections of the prefrontal cortex in the rhesus monkey. In P. Chauvel, A. V. Delgado-Escueta, E. Halgren & J. Bancaud (Eds.), *Advances in neurology, No. 57*. (pp. 91-115). New York: Raven Press.

Barbas, H. (1995). Anatomic basis of cognitive-emotional interactions in the primate prefrontal cortex. *Neuroscience and Behavioral Review*, *19*, 499-510.

Barbas, H. (2000). Connections underlying the synthesis of cognition, memory and emotion in primate prefrontal cortices. *Brain Research Bulletin*, *52*, 5, 319-330.

Barbas, H., Medalla, M., Alade, O., Suski, J., Zikopoulos, B., & Lera, P. (2005). Relationship of prefrontal connections to inhibitory systems in superior temporal areas in the rhesus monkey. *Cerebral Cortex*, *15*, 9, 1356-1370.

Barbas, H., & Pandya, D. N. (1989). Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. *Journal of Comparative Neurology*, *286*, 353–375.

Barbas, H., & Rempel-Clower, N. (1997). Cortical structure predicts the pattern of corticocortical connections. *Cerebral Cortex*, *7*, 635–646.

Barone, P., & Joseph, J. P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, *78*, 447-464.

Barraclough, D. J., Conroy, M. L. & Lee, D. (2004). Prefrontal cortex and decision-making in a mixed-strategy game. *Nature Neuroscience*, *7*, 4, 404-410.

Besner, D., & Davelaar, E. (1982). Basic processes in reading: Two phonological codes. *Canadian Journal of Psychology*, *36*, 707-711.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6,* 173-189.

Bireta, T. J., Neath, I., & Suprenant, A. M. (2006). The syllable based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review, 13*, 3, 434-438.

Blasdel, G. G. (1992a). Differential imaging of ocular dominance and orientation selectivity in monkey striate cortex. *Journal of Neuroscience, 12,* 3115-3138.

Blasdel, G. G. (1992b). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience, 12*, 3139-3161.

Boardman, I., & Bullock, D. (1991). A neural network model of serial order recall from shortterm memory. *Proceedings of the International Joint Conference on Neural Networks*, *2*, 879-884. Piscataway, NJ: IEEE Service Center.

Bor, D., Cumming, N., Scott, C. E. L., & Owen, A. M. (2004). Prefrontal cortical involvement in verbal encoding strategies. *European Journal of Neuroscience*, *19*, 3365-3370.

Bor, D., Duncan, J., Wiseman, R. J., & Owen, A. M. (2003). Encoding Strategies Dissociate Prefrontal Activity from Working Memory Demand. *Neuron*, *37*, 361–367.

Botvinick, M. M., & Plaut, D. C. (2006a). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113*, 201-233.

Botvinick, M. M., & Plaut, D. C. (2006b). Such stuff as habits are made on: A reply to Cooper and Shallice (2006). *Psychological Review, 113,* 917-928.

Bower, G. H. (1972). Perceptual groups as coding units in immediate memory. *Proceedings of the Psychonomic Society*, *27*, 4, 217-219.

Bradski, G., Carpenter, G. A., & Grossberg, S. (1992). Working memory networks for learning multiple groupings of temporal order with application to 3-D visual object recognition. *Neural Computation*, *1992, 4*, 270-286.

Bradski, G., Carpenter, G. A., & Grossberg, S. (1994). STORE working memory networks for storage and recall of arbitrary temporal sequences. *Biological Cybernetics*, *71*, 469-480.

Bradski, G., & Grossberg, S. (1995). Fast learning VIEWNET architectures for recognizing 3-D objects from multiple 2-D views. *Neural Networks, 8*, 1053-1080.

Braver, T. S., Barch, D. M., Kelley, W. M., Buckner, R. L., Cohen, N. J., Miezin, F. M., Snyder, A. Z., Ollinger, J. M., Akbudak, E., Conturo, T. E., & Petersen, S. E. (2001). Direct comparison of prefrontal cortex regions engaged by working and long-term memory tasks. *Neuroimage*, *14,* 48-59.

Briggs, F., & Callaway, E. M. (2001). Layer-specific input to distinct cell types in layer 6 of monkey primary visual cortex. *Journal of Neuroscience*, *21*, 3600-3608.

Brodie, D.A., & Murdock, B.B. Jr. (1977). Effect of presentation time on nominal and functional serial-position curves of free recall. *Journal of Verbal Learning and Verbal Behavior, 16,* 185-200.

Brodie, D. A., & Prytulak, L. S. (1975). Free recall curves: Nothing but rehearsing some items more or recalling them sooner? *Journal of Verbal Learning & Verbal Behavior*, *14*, 549-563.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*, 127-181.

Brown, J., Bullock, D., and Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience, 19,* 10502-10511.

Brown, J. W., Bullock, D., & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, *17*, 471-510.

Bullock, D. (2004). From Parallel Sequence Representations to Calligraphic Control: A Conspiracy of Neural Circuits. *Motor Control*, *8*, 371-391.

Buchanan, F.M., & Bitterman, M.E. (1998). Learning in honeybees as a function of amount and frequency of reward. *Animal Learning and Behavior, 16*, 247-255.

Bullock, D., Bongers, R., Lankhorst, M., & Beek, P. J. (1999). A vector-integration to-endpoint model for performance of viapoint movements. *Neural Networks*, *12*, 1-29.

Bullock, D., & Grossberg, S. (1988a). Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review*, *95*, 49-90.

Bullock, D., & Grossberg, S. (1988b). The VITE model: A neural command circuit for generating arm and articulatory trajectories. In J. A. S. Kelso, A. J. Mandell & M. F. Schlesinger (Eds.), *Dynamic Patterns in Complex Systems* (pp. 305-326). Singapore: World Scientific Publishers.

Bullock, D., & Grossberg, S. (1989). VITE and FLEET: neural modules for trajectory formation and postural control. In A. Hershberger (Ed.), *Volitional Action* (pp. 253-298). Amsterdam: Elsevier/North Holland.

Bullock, D., & Grossberg, S. (1991). Adaptive neural networks for control of movement invariant under speed and force rescaling. *Human Movement Science*, *10*, 1, 3-53.

Bullock, D., Grossberg, S., & Guenther, F.H. (1993). A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. *Journal of Cognitive Neuroscience*, *5*, 408-435.

Bullock, D., Grossberg, S., & Mannes, C. (1993). A neural network model for cursive script production. *Biological Cybernetics*, *70*, 15-28.

Bullock, D., & Rhodes, B. (2003). Competitive queuing for serial planning and performance. In M. Arbib (Ed.), *Handbook of brain theory and neural networks, 2ed.* (pp. 241-244). Cambridge, MA: MIT Press.

Burgess, N. (1995). A solvable connectionist model of immediate recall of ordered lists. In G. Tesauro, D.S. Touretzky & T.K. Leen (Eds.), *Advances in Neural Information Processing Systems*, *7*, Cambridge, Mass.: MIT Press.

Burgess, N., & Hitch, G. J. (1992). Toward a network model of the articulatory loop. *Journal of Memory and Language*, *31*, 429-460.

Burgess, N., & Hitch, G. J. (1996). A connectionist model of STM for serial order. In S. E. Gathercole (Ed.), *Models of short-term memory* (pp. 51-72). Hillsdale, NJ: Erlbaum.

Burgess, N., & Hitch, G. J. (1999). Memory for Serial Order: A Network Model of the Phonological Loop and its Timing. *Psychological Review*, *106*, 3, 551-581.

Bussey, T. J., Wise, S. P., & Murray, E. A. (2001). The role of ventral and orbital prefrontal cortex in conditional visuomotor learning and strategy use in rhesus monkeys (*Macaca mulatta*). *Behavioral Neuroscience*, *115*, 971-982.

Callaway, E. M. (1998). Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, *21*, 47-74.

Callaway, E. M., & Wiser, A. K. (1996). Contributions of individual layer 2-5 spiny neurons to local circuits in macaque primary visual cortex. *Visual Neuroscience*, *13*, 907-922.

Caplan, D., Rochon, E., & Waters, G. (1992). Articulatory and phonological determinants of word-length effects in span tasks. *Quarterly Journal of Experimental Psychology*, *45A*, 177–192.

Carlesimo, G. A., Marfia, G. A., Loasses, A., & Caltagirone, C. (1996). Recency effect in anterograde amnesia: Evidence for distinct memory stores underlying enhanced retrieval of terminal items in immediate and delayed recall paradigms. *Neuropsychologia, 34,* 177–184.

Carpenter G. A. & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. Cambridge, MA: MIT Press.

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.

Chafee, M. V., & Goldman-Rakic, P. S. (1998). Neuronal activity in macaque prefrontal area 8a and posterior parietal area 7ip related to memory guided saccades. *Journal of Neurophysiology*, *79*, 2919–2940.

Chafee, M. V., & Goldman-Rakic, P. S. (2000). Inactivation of Parietal and Prefrontal Cortex Reveals Interdependence of Neural Activity During Memory-Guided Saccades. *Journal of Neurophysiology*, *83*, 1550-1566.

Chein, J. M., & Feiz, J. A. (2001). Dissociation of Verbal Working Memory System Components Using a Delayed Serial Recall Task. *Cerebral Cortex*, *11*, 1003-1014.

Chey, J., Grossberg, S., & Mingolla, E. (1997). Neural dynamics of motion grouping: From apertureambiguity to object speed and direction. *Journal of the Optical Society of America, 14,* 2570-2594.

Church, R.M., & Meck, W.H. (1984). The numerical attribute of stimuli. In H. L. Roitblat, T. G. Bever, & H. J. Terrace (Eds.). *Animal cognition*. Hillsdale, NJ: Erlbaum.

Cisek, P. (2005). Neural representations of motor plans, desired trajectories, and controlled objects. *Cognitive Processes*, 6, 15–24.

Cisek, P., Grossberg, S., & Bullock, D. (1998). A cortico-spinal model of reaching and proprioception under multiple task constraints. *Journal of Cognitive Neuroscience*, *10*(4), 425-444.

Cisek, P., & Kalaska, J. F. (2002). Simultaneous encoding of multiple potential reach directions in dorsal premotor cortex. *Journal of Neurophysiology*, *87*, 1149-1154.

Clower, W. T., & Alexander, G. E. (1998). Movement sequence-related activity reflecting numerical order of components in supplementary and presupplementary motor areas. *Journal of Neurophysiology*, *80*, 1562–1566.

Cohen, M., & Grossberg, S. (1986). Neural dynamics of speech and language coding: developmental programs, perceptual grouping, and competition for short-term memory. *Human Neurobiology*, *5*, 1-22.

Cohen, M. A., & Grossberg, S. (1987). Masking Fields: A massively parallel neural architecture for learning, recognizing and predicting multiple groupings of patterned data. *Applied Optics*, *26*, 1866-1891.

Cohen, M. A., Grossberg, S., & Stork, D. G. (1988). Speech perception and production by a self-organizing neural network. In Y.C. Lee (Ed.), *Evolution, learning, cognition, and advanced architectures* (pp. 217-231). Hong Kong: World Scientific Publishers.

Constantinidis, C., & Procyk, E. (2004). The primate working memory networks. *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 4, 444-465

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Science*, *5*, 12, 539-546.

Cooper, R. P., & Shallice, T. (2000). Contention Scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*, 297–338.

Cowan, N. (1992a). Verbal Memory Span and the Timing of Spoken Recall. *Journal of Memory and Language*, *31*, 668-684.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-185.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: a logical sequel to Miller (1956). *Psychological Science*, *15*, 9, 634-640.

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992b). The Role of Verbal Output Time in the Effects of Word Length on Immediate Memory. *Journal of Memory and Language*, *31*, 1-17.

Cowan, N., Nugent, L. D., Elliott, E. M., Ponomarev, I., & Saults, J. S. (1999). The Role of Attention in the Development of Sort-Term Memory: Age Differences in the Verbal Span of Apprehension. *Child Development*, *70*, 5, 1082-1097.

Cowan, N., Saults, J. S., Elliott, E. M., & Moreno, M. V. (2002). Deconfounding Serial Recall. *Journal of Memory and Language*, *46*, 153-177.

Cowan, N., Saults, J. S., & Nugent, L. D (1997a). The role of absolute and relative amounts of time in forgetting within immediate memory: The case of tone-pitch comparisons. *Psychonomic Bulletin & Review*, *4*, 3, 393-397.

Cowan, N., Wood, N. L., Nugent, L. D., & Treisman, M. (1997b). These are two word-length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science*, *8*, 4, 290-295.

Cowan, N., Wood, N. L., Wood, P. K., Keller, T. A., Nugent, L. D., & Keller, C. V. (1998). Two separate verbal processing rates contributing to short-term memory span. *Journal of Experimental Psychology-General*, *127*, 2, 141-160.

Craik, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior, 9,* 143–148.

Crannell, C. W., & Parrish, J. M. (1957). A comparison of immediate memory span for digits, letters and words. *The Journal of Psychology, 44*, 319–327.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

Crowder, R. G. (1978). Mechanisms of auditory backward masking in the stimulus suffix effect. *Psychological Review, 85,* 502–524.

Curtis, C. E., Rao, V. Y., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Science*, *7*, 9, 415-423.

Curtis, C. E., & D'Esposito, M. (2004). The effects of prefrontal lesions on working memory performance and theory. *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 4, 528-539.

Cusick, C. G., Seltzer, B., Cola, M., & Griggs, E. (1995). Chemoarchitectonics and corticocortical terminations within the superior temporal sulcus of the rhesus monkey: evidence for subdivisions of superior temporal polysensory cortex. *Journal of Computational Neurology*, *296*, 159-172.

Dalezman, J. J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 597–608.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The Demise of Short-Term Memory Revisited: Empirical and Computational Investigations of Recency Effects. *Psychological Review*, *112*, 1, 3–42.

Davelaar, E. J., Haarmann, H. J., Goshen-Gottstein, Y., & Usher, M. (2006). Semantic similarity dissociates short- from long-term recency effects: testing a neurocomputational model of list memory. *Memory and Cognition*, *34*, 2, 323-34.

Dehaene, S. (1997). The number sense: How the mind creates mathematics. New York: Oxford University Press.

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review*, *104*, 123-147.

DeLosh, E. L., & McDaniel, M. A. (1996). The Role of Order Information in Free Recall: Application to the Word-Frequency Effect. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*, 5, 1136-1146.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Science*, *USA*, *93*, 13494-13499.

D'Esposito, M., Aguirre, G. K., Zarahn, E., Ballard, D., Shin, R. K., & Lease, J. (1998). Functional MRI studies of spatial and non-spatial working memory. *Cognitive Brain Research*, *7*, 1-13.

D'Esposito, M., & Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia*, *37*, 1303-1315.

Dessing, J. C., Caljouw, S. R., Peper, C. E., & Beek, P. J. (2004). A dynamical neural network for hitting an approaching object. *Biological Cybernetics*, *91*, 377-387.

Dessing, J. C., Peper, C. E., Bullock, D., & Beek, P. J. (2005). How Position, Velocity, and Temporal Information Combine in the Prospective Control of Catching: Data and Model. *Journal of Cognitive Neuroscience*, *17*, 668-686.

Dombrowski, S. M., Hilgetag, C. C., & Barbas, H., (2001). Quantitative architecture distinguishes prefrontal cortical systems in the rhesus monkey. *Cerebral Cortex*, *11*, 975–988.

Dosher, B. A., & Ma, J. (1998). Output loss or rehearsal loop? Output-time versus pronunciation-time limits in immediate recall for forgetting matched materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 316-335.

Drewnowski, A. (1980). Attributes and priorities in short-term recall: a new model of memory span. *Journal of Experimental Psychology: General, 109*, 208–250.

Drewnowski, A., & Murdock, B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 319–332.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature*, *384*, 159–61.

Eacott, M. J., & Gaffan, D. (1992). Inferotemporal-frontal disconnection: The uncinate fascicle and visual associative learning in monkeys. *European Journal of Neuroscience*, *4*, 1320-1332.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Emmerton, J., Lohmann, A., & Niemann, J. (1997). Pigeons' serial ordering of numerosity with visual arrays. *Animal Learning and Behavior, 25*, 234-244.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience, 2*, 704 –716.

Estes, W. K. (1972). An associative basis for coding and organization in memory. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 161-190). Washington, DC: Winston.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, *9*, 59–79.

Farrell, S., & Lewandowsky, S. (2004). Modeling transposition latencies: Constraints for theories of serial order memory. *Journal of Memory and Language*, *51*, 115-135.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Fletcher, P. C., & Henson, R. N. A. (2001). Frontal lobes and human memory: Insights from functional neuroimaging. *Brain*, *124*, 849-881.

Frey, S., Kostopoulos, P., & Petrides, M. (2004). Orbitofrontal contribution to auditory encoding. *Neuroimage*, *22*, 3, 1384-1389.

Frey, S., & Petrides, M. (2002). Orbitofrontal cortex and memory formation. *Neuron*, *36*, 1, 171-176.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*, 331-349.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1990). Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. *Journal of Neurophysiology*, *63*, 814-831.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1991). Neuronal activity related to saccadic eye movements in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *65*, 1464-1483.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1993). Dorsolateral prefrontal lesions and oculomotor delayed-response performance: Evidence for mnemonic "scotomas." *Journal of Neuroscience*, *13*, 1479-1497.

Funahashi, S., Inoue, M., & Kubota, K. (1993). Delay-related activity in the primate prefrontal cortex during sequential reaching tasks with delay. *Neuroscience Research*, *18*, 171-175.

Funahashi, S., Inoue, M., & Kubota, K. (1997). Delay-period activity in the primate prefrontal cortex encoding multiple spatial positions and their order of presentation. *Behavioral Brain Research*, *84*, 203-223.

Fuster, J. M. (1973). Unit activity in the prefrontal cortex during delayed response performance: neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*, 61-78.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*, 652-654.

Fuster, J. M., Bauer, R. H., & Jervey, J. P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research*, *330*, 299-307.

Gaudiano, P., & Grossberg, S. (1991). Vector associative maps: Unsupervised real-time error-based learning and control of movement trajectories. *Neural Networks*, *4*, 147-183.

Glanzer, M. (1972). Storage mechanisms in recall. In *The psychology of learning and motivation: Advances in research and theory, Vol. V.* (G. H. Bower, Ed.). New York: Academic Press.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5*, 351 – 360.

Goldman, P. S., & Rosvold, H. E. (1970). Localization of function within the dorsolateral prefrontal cortex of the rhesus monkey. *Experimental Neurology*, 27, 291-304.

Goldman, P. S., Rosvold, H. E., Vest, B., & Galkin, T. W. (1971). Analysis of the delayed-alternation deficit produced by dorsolateral prefrontal lesions in the rhesus monkey. *Journal of Comparative & Physiological Psychology*, 77, 212-220.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum, & V. Mountcastle (Eds.), *Handbook of Physiology*, (Vol. 5, pp. 373-417). American Physiological Society, Bethesda.

Goodglass, H., & Kaplan, E. (2001). *The Assessment of Aphasia and Related Disorders* (2nd ed.). Philadelphia: Lea and Febiger.

Grossberg, S. (1969). On the serial learning of lists. *Mathematical Biosciences, 4*, 201-253.

Grossberg, S., & Pepe, J. (1971). Spiking threshold and overarousal effects in serial learning. *Journal of Statistical Physics*, *3*, 95-125.

Grossberg, S. (1971). On the dynamics of operant conditioning. *Journal of Theoretical Biology*, *33,* 225-255.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, *52*, 213-257.

Grossberg, S. (1974). Classical and instrumental learning by neural networks. In R. Rosen and F. Snell (Eds.), *Progress in theoretical biology.* New York: Academic Press, pp. 51-141. (Reprinted in Grossberg, S. (1982). *Studies of Mind and Brain* (pp. 500-639). Boston: Reidel.)

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. (1976). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, *23,* 187-202.

Grossberg, S. (1978a). A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans. In B. Rosen & F. Snell (Eds.), *Progress in*

Theoretical Biology, (Vol. 5, pp. 233-374). New York: Academic Press. (Reprinted in Grossberg, S. (1982). *Studies of Mind and Brain* (pp. 500-639). Boston: Reidel.)

Grossberg, S. (1978b). Behavioral contrast in short-term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, *17*, 199-219.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1-51.

Grossberg, S. (1982). Processing of expected and unexpected events during conditioning and attention: A psychophysiological theory. *Psychological Review, 89*, 529-572.

Grossberg, S. (1984). Outline of a theory of brightness, color, and form perception. In E. Degreef & J. van Buggenhaut (Eds.), *Trends in mathematical psychology.* Amsterdam: North-Holland.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of perceptual grouping: Textures, boundaries and emergent segmentations. *Perception and Psychophysics*, *38*, 141–171.

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

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, *1*, 17-61.

Grossberg, S. (1999a). How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex. *Spatial Vision*, *12*, 163-185.

Grossberg, S. (1999b). The link between brain learning, attention, and consciousness. *Consciousness and Cognition*, *8*, 1-44.

Grossberg, S. (2000a). How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society*, 6, 579-588.

Grossberg, S. (2000b). The complementary brain: Unifying brain dynamics and modularity. *Trends in Cognitive Sciences, 4*, 233-246.

Grossberg, S. (2003a). How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, *2*, 47-76.

Grossberg, S. (2003b). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, 423-445.

Grossberg, S., Boardman, I., & Cohen, C. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 418-503.

Grossberg, S., & Howe, P. D. L. (2003). A laminar cortical model of stereopsis and three-dimensional surface perception. *Vision Research*, *43*, 801-829.

Grossberg, S., & Kuperstein, M. (1986). *Neural dynamics of adaptive sensory-motor control: Ballistic eye movements* (1989: *Expanded edition*). Elmsford , NY: Pergamon Press.

Grossberg, S., & Levine, D. S. (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: Blocking, inter-stimulus interval, and secondary reinforcement. *Applied Optics*, 26, 5015-5030.

Grossberg, S., & Merrill, J. W. L. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cognitive Brain Research*, *1*, 3-38.

Grossberg, S., & Merrill, J. W. L. (l996). The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience, 8*, 257-277.

Grossberg, S., Mingolla, E., & Ross, W. D. (1997). Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neuroscience*, *20*, 106-111.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: interword integration and duration-dependent backward effects. *Psychological Review*, *107*, 735–767.

Grossberg, S., & Paine, R. W. (2000). A neural model of corticocerebellar interactions during attentive imitation and predictive learning of sequential handwriting movements. *Neural Networks, 13*, 999-1046.

Grossberg, S., & Raizada, R. D. S. (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research*, *40*, 1413-1432.

Grossberg, S., & Repin, D. (2003). A neural model of how the brain represents and compares multi-digit numbers: spatial and categorical processes. *Neural Networks*, *16,* 1107-1140.

Grossberg, S. & Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, *2*, 79-102.

Grossberg, S., & Seidman, D. (2006). Neural dynamics of autistic behaviors: Cognitive, emotional, and timing substrates. *Psychological Review*, *113*, 483-525.

Grossberg, S., & Seitz, A. (2003). Laminar Development of Receptive Fields, Maps and Columns in Visual Cortex: The Coordinating Role of the Subplate. *Cerebral Cortex*, *13*(8), 852-863.

Grossberg, S., & Stone, G. O. (1986a). Neural dynamics of attention switching and temporal order information in short-term memory. *Memory and Cognition*, *14*, 451-468.

Grossberg, S., & Stone, G. O. (1986b). Neural Dynamics of Word Recognition and Recall: Attentional Priming, Learning, and Resonance. *Psychological Review*, *93*(1), 46-74.

Grossberg, S., & Swaminathan, G. (2004). A laminar cortical model for 3D perception of slanted and curved surfaces and of 2D images: development, attention and bistability. *Vision Research*, *44*, 11, 1147-1187.

Grossberg, S., & Versace, M. (2006). From synapse to self: Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. Technical Report CAS/CNS-TR-06-012.

Grossberg, G., & Williamson, J. R. (2001). A neural model of how horizontal and interlaminar connections of visual cortex develop into adult circuits that carry out perceptual groupings and learning. *Cerebral Cortex, 11*, 37-58.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*, 43-53.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594-621.

Guenther, F.H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*, 350-365.

Guenther, F. H., Ghosh, S. S., & Nieto-Castanon, A. (2003). A neural model of speech production. *Proceedings of the 6th International Seminar on Speech Production, Sydney, Australia*.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*, 611-633.

Habeck, C., Rakitin, B. C., Moeller, J., Scarmeas, N., Zarahn, E., Brown, T., & Stern, Y. (2005). An event-related fMRI study of the neural networks underlying the encoding, maintenance, and retrieval phase in a delayed-match-to-sample task. *Cognitive Brain Research*, *23*, 2-3, 207-220.

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science*, *274*, 427–430.

Hartley, T. A., & Houghton, G. (1996). A linguistically constrained model of short-term memory for nonwords. *Journal of Memory and Language*, *35*, 1-31.

Hasegawa, R. P., Blitz, A. M., & Goldberg, M. E. (2004). Neurons in monkey prefrontal cortex whose activity tracks the progress of a three-step self-ordered task. *Journal of Neurophysiology, 92*, 1524-1535.

Henry, L. A. (1991). The effects of word-length and phonemic similarity in young children's short-term memory. *The Quarterly Journal of Experimental Psychology*, *43A*, 35–52.

Henson, R. N. A. (1998a). Item repetition in short-term memory: Ranschburg repeated. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 1162-1181.

Henson, R. N. A. (1998b). Short-term memory for serial order: The start-end model of serial recall. *Cognitive Psychology*, *36*, 73-137.

Henson, R. N. A. (2001). Short-term memory for serial order. *The Psychologist*, *14*, 2, 70-73.

Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology*, *49A*, 80-115.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*, 67-99.

Hikosaka, K., & Watanabe, M. (2000). Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cerebral Cortex, 10*, 3, 263-71.

Hikosaka, O., & Wurtz, R. H. (1989). The basal ganglia. *Reviews of Oculomotor Research*, *3*, 257-281.

Hilgetag, C. C., Dombrowski, S. M., & Barbas, H. (2002). Classes and gradients of prefrontal cortical organization in the primate. *Neurocomputing*, *44-46*, 823-829.

Hillis, A. E., Work, M., Barker, P. B., Jacobs, M. A., Breese, E. L., & Maurer, K. (2004). Re-examining the brain regions crucial for orchestrating speech articulation. *Brain*, *127*, 7, 1479-1487.

Hodgkin, A. L., & Huxley, A. F. (1952). A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve. *Journal of Physiology*, *117*, 500-544.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In R. Dale, C. Mellish & M. Zock (Eds.), *Current Research in Natural Language Generation* (pp. 287-319). Academic Press, London.

Houghton, G., & Hartley, T. (1996). Parallel Models of Serial Behavior: Lashley Revisited. *PSYCHE*, *2*, 25.

Horak, F. B., & Anderson, M. E. (1984). Influence of globus pallidus on arm movements in monkeys. I. Effects of kainic acid-induced lesions. *Journal of Neurophysiology*, *52*, 290-304.

Horak, F. B., & Anderson, M. E. (1984). Influence of globus pallidus on arm movements in monkeys. II. Effects of stimulation. *Journal of Neurophysiology*, *52*. 305-322.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 923–941.

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*, 269-299.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195,* 215-243.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process

in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *23*, 5, 1217-1232.

Hulme, C., Suprenant, A. M., Bireta, T. J., Stuart, G., & Neath, I. (2004). Abolishing the Word-Length Effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1, 98-106.

Hunt, R. R., & Lamb, C. A. (2001). What causes the isolation effect? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 6, 1359-66.

Isoda, M., & Tanji, J. (2004). Participation of the primate presupplementary motor area in sequencing multiple saccades. *Journal of Neurophysiology*, *92*, 653-659.

Inoue, M., & Mikami, A. (2006). Prefrontal activity during serial probe reproduction task: encoding, mnemonic and retrieval processes. *Journal of Neurophysiology*, *95*, 1008-1041.

Jacobs, J. P., & Bullock, D. (1998). A two-process model for control of legato articulation across a wide range of tempos during piano performance. *Music Perception*, *16*, 169-199.

Jacobsen, C. F. (1936). Studies of cerebral function in primates: I. The functions of the frontal association area in monkeys. *Comparative Psychology Monographs*, *13*, 1-60.

Jarrold, C., Hewes, A. K., & Baddeley, A. D. (2000). Do two separate speech measures constrain verbal short term memory in children? *Journal of Experimental Psychology*, *26*, 6, 1626-1637.

Jensen, O., & Lisman, J. E. (1996). Novel lists of 7 +/- 2 known items can be reliably stored in an oscillatory short-term memory network: interaction with long-term memory. *Learning & Memory*, *3*, 257-263.

Johnson, G. J. (1991). A distinctiveness model of serial learning. *Psychological Review*, *98*, 204-217.

Jones, D., Farrand, P., Stuart, G., & Morris, N. (1995). The functional equivalence of verbal and spatial memory in serial short-term memory, *Journal of Experimental Psychology: Learning Memory and Cognition, 21,* 1008-1018.

Jordan, M. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the 8th Annual Conference of the Cognitive Science Society* (pp. 10-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*, 103–109.

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review, 12*, 1, 159-164.

Kahana, M. J., & Jacobs, J. (2000). Interresponse Times in Serial Recall: Effects of Intraserial Repetition. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *26*, 5, 1188-1197.

Kandel, E. R., Schwartz, J. H., & Jessel, T. M (2000). *Principles of Neural Science*. McGraw-Hill.

Kapadia, M. K, Ito, M., Gilbert, C. D., & Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron, 15*, 843–856.

Kayser, A. S., & Miller, K. D. (2002). Opponent inhibition: A developmental model of layer 4 of the neocortical circuit. *Neuron*, *33*, 131-142.

Kermadi, I., & Joseph, J. P. (1995). Activity in the caudate nucleus of monkey during spatial sequencing. *Journal of Neurophysiology*, *74*, 911-933.

Kermadi, I., Jurquiet, Y., Arzi, M., & Joseph, J. P. (1993). Neural activity in the caudate nucleus of monkeys during spatial sequencing. *Experimental Brain Research*, *94*, 352-356.

Khader, P., Burke, M., Bien, S., Ranganath, C., & Rosler, F. (2005). Content-specific activation during associative long-term memory retrieval. *Neuroimage*, *27*, 4, 805-816.

Klein, K. A., Addis, K. M., & Kahana, M. J. (2005). A comparative analysis of serial and free recall. *Memory & Cognition*, *33*, 5, 833-839.

Knoedler, A. J., Hellwig, K. A. & Neath, I. (1999). The shift from recency to primacy with increasing delay. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 474-487.

Kostopoulos, P., & Petrides, M. (2003). The mid-ventrolateral prefrontal cortex: insights into its role in memory retrieval. *European Journal of Neuroscience*, *17*, 7, 1489-1497.

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*, 337-347.

Lashley, K. (1951). The problem of serial order in behavior. In Jeffress (Ed.), *Cerebral Mechanisms in Behavior* (pp. 112-136). L. A. New York: Wiley.

Lee, C. L., & Estes, W. K. (1977) Order and position in primary memory for letter strings. *Journal of Verbal Learning and Verbal Behavior*, *16*, 395-418.

Lee, D. N. (1976). A theory of visual control of braking based on information about time-to-collision. *Perception*, *5*, 437–459.

Levy, R., & Goldman-Rakic, P. (2000). Segregation of working memory functions within the dorsolateral prefrontal cortex. *Experimental Brain Research*, *133*, 23–32.

Lewandowsky, S. L., & Murdock, B. B. Jr. (1989). Memory for Serial Order. *Psychological Review*, *96*, 1, 25-57.

Lovatt, P. J., Avons, S. E., & Masterson, J. (2000). The word-length effect and disyllabic words. *Quarterly Journal of Experimental Psychology*, *53*, 1–22.

Lovatt, P. J., Avons, S. E., & Masterson, J. (2002). Output Decay in Immediate Serial Recall: Speech Time Revisited. *Journal of Memory and Language*, *46*, 227–243.

MacNeilage, P. F. (1964). Typing errors as clues to serial ordering mechanisms in language behaviour. *Language and Speech, 7,* 144-159.

Mannes, C. (1994). Neural network models of serial order and handwriting movement generation. Ph.D. Dissertation, Department of Cognitive and Neural Systems, Boston University.

Marshuetz, C., Smith, E. E., Jonides, J., DeGutis, J., & Chenevert T. L. (2000). Order information in working memory: fMRI evidence for parietal and prefrontal mechanisms. *Journal of Cognitive Neuroscience, 12*, 2, 130-44.

Maunsell, J. H. R., & van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus duration, speed, and orientation. *Journal of Neurophysiology, 49*, 1127-1147.

Mechelli, A., Price, C. J., Friston, K. J., & Ishai, A. (2004). Where bottom-up meets top-down: Neuronal interactions during perception and imagery. *Cerebral Cortex*, *14*, 1256-1265.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375-407.

Middleton, F. A., & Strick, P. L. (1997). Dentate output channels: motor and cognitive components. In C. I. de Zeeuw, P. Strata, & J. Voogd (Eds.), *Progress in Brain Research*, *114*, 553-566. Elsevier Science.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*, 2, 81-97.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*, 5154-5167.

Mishkin, M. (1954). Visual discrimination performance following partial ablations of the temporal lobe: II. Ventral surface vs. hippocampus. *Journal of Comparative & Physiological Psychology*, *47*, 187-193.

Mishkin, M. (1957). Effects of small frontal lesions on delayed alternation in monkeys. *Journal of Neurophysiology*, *20*, 615-622.

Mishkin, M., & Manning, F. J. (1978). Non-spatial memory after selective prefrontal lesions in monkeys. *Brain Research*, *143*, 313-323.

Mishkin, M., & Pribram, K. H. (1954). Visual discrimination performance following partial ablations of the temporal lobe: I. Ventral vs. lateral. *Journal of Comparative & Physiological Psychology*, *47*, 14-20.

Mishkin, M., & Pribram, K. H. (1955). Analysis of the effects of frontal lesions in monkey: I. Variations of delayed alternation. *Journal of Comparative & Physiological Psychology*, *48*, 492-495.

Mishkin, M., & Pribram, K. H. (1956). Analysis of the effects of frontal lesions in monkey: II. Variations of delayed response. *Journal of Comparative & Physiological Psychology*, *49*, 36-40.

Mishkin, M., Vest, B., Waxler, M., & Rosvold, H. E. (1969). A reexamination of the effects of frontal lesions on object alternation. *Neuropsychologia*, *7*, 357-364.

Munoz, D., & Wurtz, R. (1995). Saccade-related activity in monkey superior colliculus. I. Characteristics of burst and buildup cells. *Journal of Neurophysiology, 73*, 2334-2348.

Murdock, B. B. (1960). The distinctiveness of stimuli. *Psychological Review*, *67*, 16-31.

Murdock, B. B. (1961). The retention of individual items. *Journal of Experimental Psychology*, *62*, 618-625.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482-488.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*, 251-269.

Narayanan, N. S., Prabhakaran, V., Bunge, S. A., Christoff, K., Fine, E. M., & Gabrieli, J.D. (2005). The role of the prefrontal cortex in the maintenance of verbal working memory: an event-related FMRI analysis. *Neuropsychology*, *19*, 2, 223-232.

Neath, I., & Nairne, J. S. (1995). Word-length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin and Review*, *2*, 429-441.

Nieder, A., & Miller, E.K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. Neuron*, 37*, 149-157.

Nieder, A., & Miller, E.K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences, 101*, 7457-7462.

Nichols, E. A., Kao, Y. C., Verfaellie, M., & Gabrieli, J. D. (2006). Working memory and long-term memory for faces: Evidence from fMRI and global amnesia for involvement of the medial temporal lobes. *Hippocampus*, *16*, 7, 604-16.

Nigrin, A. L. (1993). *Neural networks for pattern recognition*. Cambridge, MA: MIT Press.

Niki, H. (1974). Prefrontal unit activity during delayed alternation in the monkey. I. Relation to direction of response. *Brain Research*, *68*, 185-196.

Ninokura, Y., Mushiake, H., & Tanji, J. (2003). Representation of the temporal order of visual objects in the primate lateral prefrontal cortex. *Journal of Neurophysiology*, *89*, 2868-2873.

Ninokura, Y., Mushiaske, H., & Tanji, J. (2004). Integration of temporal order and object information in the monkey lateral prefrontal cortex. *Journal of Neurophysiology*, *91*, 555-560.

Nipher, F. E. (1878). On the distribution of errors in numbers written from memory. *Transactions of the Academy of Science of St. Louis*, *3*, ccx–ccxi.

Norris, D., Page, M. P. A., & Baddeley, A. D. (1994). *Serial recall: It's all in the representations*. Paper presented at the International Conference on Working Memory, Cambridge, England.

Olson, S., & Grossberg, S. (1998). A neural network model for the development of simple and complex cell receptive fields within cortical maps of orientation and ocular dominance. *Neural Networks*, *11*, 189-208.

Owen, A. M., Evans, A. C., & Petrides, M. (1996). Evidence for a two-stage model of spatial working memory processing within the lateral frontal cortex: a positron emission tomography study. *Cerebral Cortex*, *6*, 1, 31-38.

Owen, A. M., Stern, C. E., Look, R. B., Tracy, I., Rosen, B. R., & Petrides, M. (1998). Functional organization of spatial and non-spatial working memory processing within the human lateral frontal cortex. *Proceedings of the National Academy of Sciences, USA, Neurobiology*, *95*, 7721–7726.

Page, M. P. A. (1994). Modelling the perception of musical sequences with self-organizing neural networks. *Connection Science*, *6*, 223-246.

Page, M. P. A. (1999). Modeling the perception of musical sequences with self-organizing neural networks. In N. Griffith & P.M. Todd (Eds.), *Musical networks* (pp. 175-198). Cambridge, MA: MIT Press.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 4, 761-781.

Paine, R. W., Grossberg, S., and Van Gemmert, A. W. A. (2004). A quantitative evaluation of the AVITEWRITE model of handwriting learning . *Human Movement Science*, *23*, 837-860.

Parker, A., & Gaffan, D. (1998). Memory after frontal/temporal disconnection in monkeys: Conditional and nonconditional tasks, unilateral and bilateral frontal lesions. *Neuropsychologia*, *36*, 259-271.

Passingham, R. (1975). Delayed matching after selective prefrontal lesions in monkeys (*Macaca mulatta*). *Brain Research*, *92*, 89-102.

Passingham, R. (1993). *The Frontal Lobes and Voluntary Action.* London: Oxford University Press.

Paulesu, E., Frith, C. D., & Frackowiak, R, S. (1993). The neural correlates of the verbal component of working memory. *Nature*, *362*, 342–5.

Pesenti, M., Thioux, M. Seron, X., & De Volder, A. (2000). Neuroanatomical substrates of Arabic number processing, numerical comparison, and simple addition: a PT study. *Journal of Cognitive Neuroscience, 12*, 461-479.

Petrides, M. (1991). Functional specialization within the dorsolateral prefrontal cortex for serial order memory. *Proceedings of the Royal Society of London Series B*, *246*, 1317, 299-306.

Petrides, M. (1994b). Frontal lobes and working memory: evidence from investigations of the effects of cortical excisions in nonhuman primates. In F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology* (Vol. 9, pp. 59–82). Amsterdam: Elsevier.

Petrides, M. (1995). Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. *Journal of Neuroscience*, *15*, 359-375.

Petrides, M. (2000a). Dissociable roles of mid-dorsolateral prefrontal and anterior inferotemporal cortex in visual working memory. *Journal of Neuroscience*, *20*, 7496-7503.

Petrides, M. (2000b). Mapping Prefrontal Cortical Systems for the Control of Cognition. In *Brain Maping: The Systems* (pp. 159-176). Academic Press.

Petrides, M. (2000c). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research*, *133*, 44-54.

Petrides, M. (2005). Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Society of London B Biological Science*, *360*, 1456, 781-795.

Petrides, M., Alivisatos, B., & Frey, S. (2002). Differential activation of the human orbital, mid-ventrolateral, and mid-dorsolateral prefrontal cortex during the processing of visual stimuli. *Proceedings of the National Academy of Sciences, USA*, *99*, 8, 5649-5654.

Petrides, M., & Pandya, D. N. (1994). Comparative architectonic analysis of the human and the macaque frontal cortex. In F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology* (Vol. 9, pp. 17–58). Amsterdam: Elsevier.

Pinel. P., Le Clec, G., van de Moortele, P.-F., Le Bihan, D., & Dehaene, S. (1999). Event-related fMRI analysis of the cerebral circuit for number comparison. *NeuroReport, 10*, 1473-1479.

Poeppel, D. (1996). A Critical Review of PET Studies of Phonological Processing. *Brain and Language*, *55*, 317-351.

Poggio, G. F. (1972). Spatial properties of neurons in striate cortex of unanesthetized macaque monkey. *Investigative Ophthalmology*, *11,* 369-377.

Poggio, G. F., & Fischer, B. (1977). Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Journal of Neurophysiology*, *40*, 1392-1405.

Poggio, G. F. & Talbot, W. H. (1981). Mechanisms of static and dynamic stereopsis in foveal cortex of the rhesus monkey. *Journal of Physiology, 315*, 469-492.

Pollen, D. A. (1999). On the neural correlates of visual perception. *Cerebral Corte*x, *9*, 4–19.

Postle, B. R., Berger, J. S., & Esposito, M. (1999). Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *Proceedings of the National Academy of Sciences*, *USA, 96*, 12959–12964.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology, 17,* 132 – 138.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search in associative memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory. Vol. 14* (pp. 207-262). New York: Academic Press.

Rainer, G., Assad, W. F., & Miller, E. K. (1998). Selective representation of relevant information within the primate prefrontal cortex. *Nature*, *393*, 577-579.

Rainer, G., Rao, S. C., & Miller, E. K. (1999). Prospective coding for objects in primate prefrontal cortex. *Philosophical Transactions of the Society of London B*, *351*, 1455-1462.

Raizada, R. D. S., & Grossberg, S. (2001). Context-sensitive binding by the laminar circuits of V1 and V2: a unified model of perceptual grouping, attention, and orientation contrast. *Visual Cognition*, *8*, 431–466.

Raizada, R. D. S., & Grossberg, S. (2003). Towards a Theory of the Laminar Architecture of Cerebral Cortex: Computational Clues from the Visual System. *Cerebral Cortex*, *13*, 1, 100-113.

Ranganath, C., & Blumenfeld, R. S. (2005). Doubts about double dissociations between short- and long-term memory. *Trends in Cognitive Science*, *9*(8), 374-380.

Ranganath, C., DeGutis, J., & D'Esposito, M. (2004). Category-specific modulation of inferior temporal activity during working memory encoding and maintenance. *Cognitive Brain Research*, *20*, 1, 37-45.

Ranganath, C., & D'Esposito, M. (2005). Directing the mind's eye: Prefrontal, inferior and medial temporal mechanisms for visual working memory. *Current Opinion in Neurobiology*, *15*, 175-182.

Ranganath, C., Johnson, M. K., & D'Esposito, M. D. (2003). Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia*, *41*, 378-389.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, *276*, 5313, 821-824.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review*, *88*, 552–572.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 163-178.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, *89*, 4, 1009-1023.

Rempel-Clower, N. L., & Barbas, H. (2000). The laminar pattern of connections between prefrontal and anterior temporal cortices in the rhesus monkey is related to cortical structure and function. *Cerebral Cortex*, *10*, 851–865.

Rhodes, B. J. (2000). *Learning-driven Changes in the Temporal Characteristics of Serial Movement Performance: A Model Based on Cortico-Cerebellar Cooperation*. Unpublished doctoral dissertation, Department of Cognitive and Neural Systems, Boston University, Boston.

Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. P. A. (2004). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. *Human Movement Science*, *23*, 683-730.

Rickard, T.C., Romero, S.G., Basso, G., Wharton, C., Flitman, S., & Grafman, J. (2000). The calculating brain: an fMRI study. *Neuropsychologia, 38*, 325-335.

Rizzolatti, G., Riggio, L., Dascola, I., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*, 31–40.

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex, 10,* 284-294.

Rolls, E. T. (2004). Convergence of sensory systems in the orbitofrontal cortex in primates and brain design for emotion. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology*, *281A*, 1, 1212-1225.

Rolls, E. T., Browning, A. S., Inoue, K., & Hernadi, I. (2005). Novel visual stimuli activate a population of neurons in the primate orbitofrontal cortex. *Neurobiology of Learning and Memory*, *84*, 2, 111-23.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60-94.

Rumelhart, D. E., & Zipser, D. (1986). Feature Discovery by Competitive Learning. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing, 1,* 152-193. London: MIT Press.

Ryan, J. (1969). Grouping and short-term memory: different means and patterns of grouping. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *52A*, 367-394.

Rypma, B., Berger, J. S., & D'Esposito, M. (2002). The influence of working-memory demand and subject performance on prefrontal cortical activity. *Journal of Cognitive Neuroscience*, *14*, 5, 721-731.

Schoenbaum, G., & Setlow, B. (2001). Integrating orbitofrontal cortex into prefrontal theory: Common processing themes across species and subdivisions. *Learning and Memory*, *8*, 134-147.

Schoenbaum, G., & Roesch, M. (2005). Orbitofrontal Cortex, Associative Learning, and Expectancies. *Neuron*, *47*, 633–636.

Schweickert, R., & Boruff, B. (1986). Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 419-425.

Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology*, *51A*, 283–304.

Service, E. (2000). Phonological complexity and word duration in immediate recall: Different paradigms answer different questions—A comment on Cowan, Nugent, Elliott, and Geer. *Quarterly Journal of Experimental Psychology*, *53A*, 661–665.

Shima, K., & Tanji, J. (1998). Both supplementary and presupplementary motor areas are crucial for the temporal organization of multiple movements. *Journal of Neurophysiology*, *80*, 6, 3247-3260.

Shima, K., & Tanji, J. (2000). Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology, 84*, 2148–2160.

Shima, K., Isoda, M., Mushiake, H., & Tanji, J. (2007). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, *445*, 315-318.

Smith, E. E., Jonides, J., Marshuetz, C., & Koeppe, R. A. (1998). Components of verbal working memory: Evidence from neuroimaging. *Proceedings of the National Academy of Sciences, USA*, *95*, 876–882.

Smith, E. L., Chino, Y., Ni, J., & Cheng, H. (1997). Binocular combination of contrast signals by striate cortical neurons in the monkey. *Journal of Neurophysiology, 78,* 366-382.

Song, J. H., & Jiang, Y. (2005). Visual working memory for simple and complex features: An fMRI study. *Neuroimage*, In Press.

Starkey, P., Spelke, E.S., & Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. *Science, 222*, 179-181.

Stern, C. E., Sherman, S. J., Kirchhoff, B. A., & Hasselmo, M. E. (2001). Medial temporal lobe activity associated with active maintenance of novel information. *Neuron*, *31*, 865-873.

Sternberg, S., Monsell, S., Knoll, R.L., & Wright, C.E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. Stelmach (Ed.),

Information processing in motor control and learning (pp.117-152). New York: Academic Press. (Reprinted in R. A. Cole, (Ed.), *Perception and Production of Fluent Speech* (pp. 469-505), 1980, Hillsdale, NJ: Erlbaum).

Tan, L., & Ward, J. A. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 6, 1589-625.

Tzeng, O. J. L. (1973). Positive recency effect in delayed free recall. *Journal of Verbal Learning & Verbal Behavior, 12,* 436-439.

Ungerleider, L. G., & Mishkin, M. (1982). *Two cortical visual systems*. In D. G. Ingle, M. A. Goodale & R. J. Q. Mansfield (Eds.), Analysis of visual behavior. Cambridge, MA: MIT Press.

Usher, M., & McClelland, J. L. (2001). On the Time Course of Perceptual choice: The leaky competing accumulator model. *Psychological Review, 108,* 550-592.

Von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science, 224,* 1260–1262.

Von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld (The effects of field formation in the trace field). *Psychologie Forschung*, *18*, 299-34.

Wang, M., Zhang, H., & Li, B. M. (2000). Deficit in conditional visuomotor learning by local infusion of bicuculline into the ventral prefrontal cortex in monkeys. *European Journal of Neuroscience*, *12*, 3787-3796.

Ward, N. (1994). *A connectionist language generator*. Norwood, NJ: Ablex Publishing.

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, *30*, 6, 885-92.

Washburn, D.A., & Rumbaugh, D.M. (1991). Ordinal judgments of numerical symbols by macaques (*Macaca mulatta*). *Psychological Science, 2*, 190-193.

Watkins, M. J., Neath, I., & Sechler, E. S. (1989). Recency effect in recall of a word list when an immediate memory task is performed after each word presentation. *American Journal of Psychology*, *102*, 2, 265-270.

Watkins, O. C., & Watkins, M. J. (1980). The Modality Effect and Echoic Persistence. *Journal of Experimental Psychology: General, 109*, 251–278.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*, 89-104.

Whitten, W. B. (1978). Output interference and long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 685–692.

Woodward, T. S., Cairo, T. A., Ruff, C. C., Takane, Y., Hunter, M. A., & Ngan, E. T. C. (2005). Functional connectivity reveals load dependent neural systems underlying encoding and maintenance in verbal working memory. *Neuroscience*, In Press.

Zhang, G., & Simon, H. A. (1985). STM capacity for Chinese words and idioms: Chunking and articulatory loop hypothesis. *Memory & Cognition*, *13*, 193-201.

APPENDIX

A.1 Cell Membrane Equations. Model cell dynamics obey membrane, or shunting, equations that interact (Hodgkin & Huxley, 1952; Grossberg, 1973). Each model cell possesses a single voltage compartment with membrane potential, *V(t)*, given by an equation of the form:

$$C_m \frac{d}{dt} V(t) = -\left(V(t) - E_{leak}\right)\gamma_{leak} - \left(V(t) - E_{inhib}\right)\gamma_{inhib}(t) - \left(V(t) - E_{excit}\right)\gamma_{excit}(t). \qquad (1)$$

In (1), the time-varying conductances $\gamma_{excit}(t)$ and $\gamma_{inhib}(t)$ represent, respectively, the total excitatory and inhibitory inputs, as described by the model architecture (Figures 2, 10, and 14), in addition to a constant leakage conductance, γ_{leak}, which maintains a resting membrane potential in the absence of inputs. Reversal potentials are set to: $E_{excit} = 1$, $E_{inhib} = -1$, $E_{leak} = 0$ except where otherwise indicated. These continuous-time differential equations were numerically integrated using a fourth-order Runge-Kutta algorithm and implemented in C++.

A.2 Simulation Conventions. Table 3 summarizes the following simulation procedures. Inputs to the cognitive working memory system, I_n, were pulse inputs of magnitude .1 and one simulation time duration. For all tasks, one simulation time unit was equivalent to 100ms. Thus, items were presented every five simulation time units to correspond to the 500ms IOIs (Figure 5) used in all serial recall tasks, and every 12.5 simulation time units to correspond to the 1.25 second IOIs used in the free recall task. The recall cue in the ISR and IFR tasks (Figures 6, 7, 8a and 9) occurred one IOI after the onset of the last item presentation. That is, the recall cue occurs at the time the expected next item fails to occur which, even in the absence of an explicit recall cue, indicates that list presentation has ceased and recall can proceed. In the delayed serial recall task (Figure 8c), the same time conversion (one simulation time unit = 100ms) was used to compute the appropriate delays for each of the retention intervals (Figure 8b) before presentation of the recall cue. For all cognitive data error simulations (Figures 6, 7, 8a, and 9), at five simulation time units after the time of the recall cue, zero mean Gaussian noise with standard deviation .015 was applied to all cognitive working memory item activities Y_i. Five simulation time units approximates the time that cognitive working memory input is gated off from motor working memory at the beginning of motor execution; see Section A.6. This noise simulation was selected to simply capture the effects of internal system noise. Similar results would be expected if the noise were applied continuously during pattern storage. Also for simplicity, no noise was applied in the motor working memory. Thus, any noise-dependent distortions of output order were inherited by the motor working memory from the cognitive working memory. The output signals, $f_5(Y_i)$ (see Equation 13), from cognitive working memory to motor working memory in Equation (10) includes a threshold (.165) that determines whether the corresponding activities are omitted or recalled.

For the free recall simulations, a re-analysis of the Tan & Ward (2000; see Figure 9b) data showed that the likelihood of rehearsing a previously presented item during a given rehearsal interval was load-dependent; that is, subjects were far less likely to rehearse items during rehearsal periods near the end of a list, when more items had been presented. Thus, at the designated time of a potential rehearsal halfway between each item presentation (6.25 simulation time units after the most recent item presentation onset), it was first determined whether any rehearsal at all would take place during this rehearsal period. If a rehearsal was chosen to occur, the item to be rehearsed was algorithmically determined by repeatedly adding a uniformly distributed random amount of activation (between 0 and .2) to each item in a parallel selection stage. These selection items start with the same activations as the current item working memory (Y_i) activations, and the activation additions continue until the item with the largest activation

level in the selection stages exceeds a fixed threshold (.5) was chosen for rehearsal. Rehearsal consisted of a re-presentation (an input of .1 magnitude and duration of one stimulus time unit) of this selected item to the working memory circuit. In delayed free recall (dotted curve in Figure 9g), no rehearsal was allowed during a 12.5 second post-list distracter period before the recall cue. In continuous-distracter free recall (dashed-dotted curve in Figure 9g), rehearsal was not allowed at all; that is, during either the ISIs or the post-list delay.

A.3 Cognitive Working Memory Storage of Activity Gradients. The cognitive working memory consists of a recurrent on-center off-surround shunting network that includes two layers of interacting cells (layers 6 and 4) that store an activity gradient.

Layer 4. The activity X_i of the ith layer 4 cell obeys the shunting on-center off-surround equation:

$$\frac{d}{dt}X_i = -.1X_i + (1 - X_i)(I_i + e[Y_i]^+) - FX_i \sum_{k \neq i} (I_k + e[Y_k]^+) \,. \tag{2}$$

As in equation (1), the three terms on the right hand side of (2) are the passive decay, excitatory, and inhibitory terms, respectively. Excitatory inputs are the bottom-up inputs, I_i, from the ith item category and from positive activation of the ith layer 6 cell output signal $[Y_i]^+$. Off-surround inhibitory inputs come from all the other bottom-up input channels ($k \neq i$) from item categories, I_k, and layer 6 output signals $[Y_k]^+$. Parameter F measures the effect of volitional gain control, V_c; see Figures 2 and 11a. For human cognitive data simulations, e = .05 and F = 1.25; for monkey neurophysiological data simulations, e = .2 and F = 15. In the simulation of continuous-distracter free recall (dashed-dotted curve in Figure 9g), F was alternated to a value of *2* (representing reduced volitional gain for storage) during the 1.15 second (1.25 second ISIs – 100ms item presentations) distracter periods following each item presentation. These transient changes in volitional gain simulate one effect of switching attention back and forth to meet the demands of the distracter task as well as of the presence of distracters per se.

Layer 6. The activity Y_i of the ith layer 6 cell obeys a shunting equation:

$$\frac{d}{dt}Y_i = -.1Y_i + (1 - Y_i)(I_i + b[X_i]^+ + \sum_j f_1(C_j)M_{ji}) \,, \tag{3}$$

where the sigmoid signal function obeys:

$$f_1(C_j) = \frac{C_j^{\,3}}{(.000001)^3 + C_j^{\,3}} \,. \tag{4}$$

Equation (3) contains only passive decay and excitatory terms. Excitatory inputs to the ith layer 6 cell are the bottom-up inputs, I_i, from the ith item category, top-down intracortical feedback from the positive layer 4 output signal $[X_i]^+$, and top-down intracortical feedback, $f_1(C_j)$, from layer 2/3 cells. These signals are multiplicatively gated by the adaptive weights, or long-term memory traces, M_{ji} that enable list chunks to read-out into the cognitive working memory the pattern of item categories that they represent. For human cognitive data: b = .7; for monkey neurophysiological data: b = 2.3. In all simulated tasks (Figures 6-9), the sequences are novel. The feedback from layer 2/3 chunk cells is therefore assumed to be negligible in the simulations, and the term is omitted.

A.4 List Chunks and Masking Field Networks

Layer 2/3. Item sequences that are stored in the cognitive working memory are grouped, or categorized, through learning into list chunk cells in layer 2/3 by a *masking field* network. A masking field is a self-similar, multiple-scale, recurrent on-center off-surround network (Cohen

& Grossberg, 1987; Grossberg, 1978a; Grossberg & Myers, 2000**).** The masking field equations for list chunks, C_j, are:

$$\frac{d}{dt}C_j = -.1C_j + (1-C_j)\left[\frac{20}{10+J}\sum_i [X_i]^+ W_{ij} + .001(J)f_1(C_j)\right] - 2[1+C_j]\frac{\sum_k f_1(C_k)K(1+(K\wedge J))}{\sum_k K(1+(K\wedge J))}, \tag{5}$$

where $f_1(C_j)$ is defined by (4). The masking field contains passive, excitatory, and inhibitory inputs. These inputs obey a *self-similarity* property whereby larger scales have more weight in the cooperative-competitive struggle for cell activation. The excitatory input consists of two terms: (1) Positive inputs, $[X_i]^+$, from layer 4 that are normalized by *1/(10 + J)*, which increases with the number *J* of inputs converging on list chunk *j,* and filtered by learned bottom-up adaptive weights, or long-term memory traces, W_{ij}, that enable the list chunk to be selectively activated due to learning; and (2) self-excitatory feedback sent through a sigmoid signal function, $f_1(C_j)$, that helps to store list chunk activity in working memory and whose strength increases with the number *J* of cortical inputs that the cell receives.

Inhibitory inputs, $\frac{\sum_k f_1(C_k)K(1+(K\wedge J))}{\sum_k K(1+(K\wedge J))}$, balance the excitatory ones. Layer 2/3 cells have different sizes and interactions whose strengths scale with cell size. Terms *J* and *K* refer to the number of inputs that cells C_j and C_k receive, respectively, and $K\wedge J$ refers to the number of inputs that the two cells C_j and C_k share. Thus, larger cells *K* inhibit a neighboring cell *J* more than smaller cells, and the strength of competition depends upon how many working memory items $K\wedge J$ are shared by *K* and *J*. The sum, $\sum K(1 + K\wedge J)$, in the denominator normalizes the total strength of inhibitory connections to each cell to equal 1. For simplicity, each possible sequence of items in cognitive working memory was represented by one list chunk.

A.5 Bottom-up and Top-Down Cognitive Learning.

Layer 6-to-2/3 Adaptive Filter. The bottom-up adaptive weight, W_{ij}, from the ith item in layer 4 to the jth list chunk in layer 2/3 is defined by the self-normalizing instar learning equation:

$$\frac{d}{dt}W_{ij} = a_1 f_2(C_j)\left[(1-W_{ij})d_1X_i - h_1W_{ij}\sum_{k\neq i}X_k\right], \tag{6}$$

where the sigmoid signal function obeys:

$$f_2(Z) = \frac{Z^3}{(.00001)^3 + Z^3}. \tag{7}$$

(Carpenter & Grossberg, 1987). In (6), activity of the list chunk C_j opens a learning gate $f_2(C_j)$ that enables weight W_{ij} to track X_i. Due to the excitatory term, $(1-W_{ij})d_1X_i$, each item *i* attempts to code a proportion of the total weight 1. The inhibitory term $h_1W_{ij}\sum_{k\neq i}X_k$ ensures that these learned weights are competitively distributed among all the items that succeed in activating C_j. Biologically, the connections between cells in layer 4 and layer 2/3 are assumed to be widely distributed, with each C_j receiving connections from a limited subset of possible X_is. For the simulation in Figure 12, however, connections were chosen such that each of the C_j cells was

connected to a different size subset of the items in the presented sequence. All weights were initially taken to have the same small initial value (.001). In all simulations, $a_1 = 1$, $d_1 = 1$, and $h_1 = 2$.

Layer 2/3-to-6 Top-Down Expectations. The top-down weight, M_{ij}, from the jth list chunk in layer 2/3 to the ith item in layer 6 is defined by the outstar learning equation:

$$\frac{d}{dt}M_{ji} = a_2 f_2(C_j)\left[Y_i - M_{ji}\right]. \tag{8}$$

(Grossberg 1968, 1980). In (8), activity of the list chunk C_j opens a learning gate, $f_2(C_j)$, that enables weight M_{ji} to track Y_i. This connectivity completes a functional loop within the cognitive working memory, in which a stored pattern in working memory activates list category chunks with activities C_j that code for that pattern through the learned layer 4-to-2/3 adaptive filter and send top-down learned expectations back to working memory through layer 2/3-to-6 connections. Biologically, these connections are presumed to be widely distributed, with each C_j sending connections to a large subset of the Y_i. For simplicity, the connections from each C_j were assumed to connect with each Y_i and to start with a small initial adaptive weight, $M_{ji} = .001$. In all simulations, $a_2 = 1$.

Layer 5 interneurons are form part of the feedback pathway (Figures 2 and 11; and are known to exist in visuo-cortical areas: Briggs & Callaway, 2001; Callaway & Wiser, 1996), but were omitted from the simulations in Figure 12 since layer 5 acts primarily as a relay from layer 2/3-to-6 in the model.

A.6 Motor Working Memory

Working Memory Storage and Selection of Motor Plans. The motor working memory consists of a network with two layers of interacting cells, a motor plan field (F_i) that stores motor plans as an activation gradient and a plan selection field (S_i) that selects the most active motor plan and forwards it to the trajectory generator. The motor plan selection circuit implements a self-normalizing variation of a competitive-queuing architecture (Grossberg, 1978a; Hartley and Houghton, 1996).

Motor Plan Field. The activity F_i of the ith motor plan field cell obeys the recurrent competitive field equation:

$$\begin{aligned} 2\frac{d}{dt}F_i = &-.1F_i + (1-F_i)\left[(1+G)f_3(F_i) + E_i + (1-f_4(G))f_5(Y_i)\right] \\ &- F_i\left[100\left[S_i - .5\right]^+ + (1+G)\sum_{k\neq i} f_3(Q_k) + \sum_{k\neq i} E_k + (1-f_4(G))\sum_{k\neq i} f_5(Y_k)\right], \end{aligned} \tag{9}$$

where the signal functions obey:

$$f_3(Z) = \frac{Z^{1.2}}{.8^{1.2} + Z^{1.2}}, \tag{10}$$

$$f_4(Z) = \frac{Z^2}{.02^2 + Z^2}, \tag{11}$$

and

$$f_5(Z) = \left[Z - .165\right]^+. \tag{12}$$

The three terms on the right hand side of (9) are the passive decay, excitatory, and inhibitory terms, respectively. The excitatory inputs are a contrast-enhancing self-excitatory feedback term, $f_3(F_i)$, boosted by the presence of a GO signal (G), a selective boost (e.g., in the monkey copying

task; see Figure 10) of visual attention, E_i, associated with eye movement and a thresholded bottom-up input, $f_5(Y_i)$, from the ith layer 6 cell in the cognitive working memory system, which is gated off by the presence of the volitional GO signal, $(1\text{-}f_4(G))$. The inhibitory inputs are derived from the excitatory inputs in the form of off-surround signals that balance the excitatory input and prevent saturation of cell activities (Grossberg, 1973). A strong specific inhibitory feedback signal, $100[S_i\text{-}.5]^+$, from supra-threshold activation of the ith plan in the plan selection field, S_i, prevents perseveration by inhibiting an item from the motor plan field after it is selected for motor output ("inhibition of return"). Off-surround ($k \neq i$) recurrent inhibitory signals, $\sum_{k \neq i} f_3(Q_k)$, come from other plans whose activity is processed by inhibitory interneurons with activities, Q_k, before being contrast-enhanced via signal, $f_3(Q_k)$. As for the excitatory on-center input, the inhibitory inputs are boosted by the GO signal. The other inputs are similar, namely the eye movement related attentional boost signals, E_k, and bottom-up inputs, $\sum_{k \neq i} f_5(Y_k)$, from layer 6 cells of cognitive working memory.

The attentional boost signals, E_k, are assumed to be pulses of duration 1 and magnitude .4 and occur two simulation time steps before item selection (in the plan selection field, S_i) during the monkey copying task. These signals are assumed to be received from a similar motor working memory that controls eye movements operating in parallel to the arm movement control system (e.g., from FEF or SEF), but for simplicity, such a parallel system is not modeled herein. They contribute to the enhancement of chosen movement commands seen in the data. The recurrent competition between items in the motor plan field is processed through inhibitory interneurons with activities, Q_i, that time-average the motor plan field activities, F_i:

$$\frac{d}{dt} Q_i = F_i - Q_i \,. \tag{13}$$

Time-averaging imposes a slight lag in inhibition and leads to smoother activity profiles in the motor plan field.

Plan Selection Field. The activity S_i of the ith plan selection field cell obeys the winner-take-all recurrent competitive field equation:

$$\frac{d}{dt} S_i = -.1 S_i + (1 - S_i)([R]^+ F_i + [S_i - .5]^+) - 20 S_i \sum_{k \neq i} [S_k - .5]^+ \,. \tag{14}$$

The three terms on the right hand side of (14) are the passive decay, excitatory, and inhibitory terms, respectively. The excitatory inputs, $([R]^+ F_i + [S_i - .5]^+)$, are a bottom-up input, F_i, from the ith cell in the motor plan field gated on by a positive rehearsal signal, $[R]^+$, and a self-excitatory feedback term when the cell exceeds an activity threshold, $[S_i - .5]^+$. Strong off-surround inhibitory signals, $20 S_i \sum_{k \neq i} [S_k - .5]^+$, come from cells in other plan selection channels ($k \neq i$). The threshold-linear self-excitatory feedback, together with threshold-linear competitive inhibition of other cells when a cell exceeds that threshold, leads to winner-take-all dynamics (Grossberg, 1973, 1980). The winning cell activates the trajectory generator through output to T_i in equation (15), and sends self-inhibitory feedback that suppresses its bottom-up input, F_i in (9), thereby preventing perseveration when the rehearsal signal again gates on bottom-up inputs, F_i, for the selection of the next motor plan.

A.7 VITE Trajectory Generator. The motor Target Position (T), Position Difference (D), Outflow Velocity ($[D]^+G$), and Present Position (P) fields correspond to the Target Position Vector (TPV), Difference Vector (DV), Outflow Velocity ($[\text{DV}]^+\text{GO}$), and Present Position

Vector (PPV), respectively, of the Vector-Integration-To-Endpoint, or VITE, model (Bullock & Grossberg, 1988a, 1991; Cisek et al., 1998).

Target Position. The activity T_i of the ith target motor position cell obeys the network equation:

$$\frac{dT_i}{dt} = -.5T_i + (1 - T_i)(100[S_i - .5]^+) . \tag{15}$$

The two terms on the right-hand side of (15) represent passive decay and excitatory terms, respectively. Equation (15) simplifies the idea that $[S_i - .5]^+$ activates a learned target position. Passive decay via term, $-.5T_i$, quickly removes the target when it is no longer supported by activity in the plan selection field. See Gaudiano and Grossberg (1991) for an analysis of how a TPV can be learned.

Position Difference. The activity D_i of the ith motor position difference cell obeys the following equation:

$$\frac{dD_i}{dt} = -D_i + [T_i]^+ - P_i . \tag{16}$$

The three terms on the right-hand side of (16) represent a passive decay, excitatory, and inhibitory terms, respectively. This cell receives excitatory input from a positive target motor position, $[T_i]^+$, and inhibitory input representing the current motor position, $-P_i$. Thus, the Motor Position Difference field (D) serves as an internal estimation of the distance between the current motor position (P) and the target motor configuration (T).

Outflow Velocity. The activity of the Position Difference field, $[D_i]^+$, is gated by the volitional GO signal, G. This gated signal, $[D_i]^+G$, scales the rate at which the Present Position, P_i, approaches the Target Position, T_i, as well as driving the differential integration-rate velocity cells, A and B, which in turn gate the selection of the next movement.

Present Position. The activity P_i of the ith present position cell obeys the following equation:

$$\frac{dP_i}{dt} = [D_i]^+ G . \tag{17}$$

Here, P_i, integrates the rectified difference, $[D_i]^+$, between present and target positions at a rate, $[D_i]^+ G$, that is determined by the volitional GO signal, G.

A.8 Motor Volition

Volitional Signal. A volitional signal, V_m, coordinates all the performance read-out processes and scales the desired rehearsal rate with the rate at which movement occurs. The volitional signal serves as excitatory input to the GO signal, G, and rehearsal signal, R. For all tasks, $V_m = 0$ during the presentation phase of the task, and 1 during the recall phases of the tasks, as indicated by the recall cue.

GO Signal. The activity, G, of the GO signal cells obey the leaky integrator equation:

$$\frac{dG}{dt} = -G + V_m . \tag{18}$$

The two terms on the right hand side of (18) are passive decay and excitatory terms, respectively. The GO signal (G) time-averages the excitatory input it receives from the volitional signal, V_m, making it positive when recall is desired and scaled by the desired performance speed. GO signal cells modulate the speed of trajectory generation via the present position cells, P in (17), and close the cognitive-to-motor read-out gate from the cognitive working memory system, Y, to the motor plan field, F, during rehearsal (see equation (9)). These signals also modulate the rate of plan selection during rehearsal, as noted below.

A.9 Variable-Rate Rehearsal

Performance Rate Estimators. The activities, A and B, of the fast and slow performance rate estimator cell populations obey the leaky integrator equations:

$$\frac{dA}{dt} = \phi\left[-.1A + G\sum_i [D_i]^+\right], \tag{19}$$

and

$$\frac{dB}{dt} = \theta\left[-.1B + G\sum_i [D_i]^+\right], \tag{20}$$

respectively. The two terms on the right hand sides of (19) and (20) are passive decay and excitatory inputs, respectively. These cells are excited by the same outflow velocity signals as the speed-sensitive cells (see (17)) and time-average these inputs at fast, ϕ, and slow, θ, rates. As a result, A and B lag the current outflow velocity signals by different degrees, so that their difference, B - A, approximately indicates the phase (increasing or decreasing portion) of a bell-shaped velocity performance curve (see Figure 16). Integration rates of the cell activities A and B (parameters ϕ and θ in equations (19) and (20), respectively) were selected so that items were selected at a rate of approximately 500ms each (consistent with the rate of presentations in the ISR simulations). For all tasks: $\phi = 3$; $\theta = 1$.

Rehearsal Signal. The activity, R, of the rehearsal signal cells obey the following equation:

$$\frac{dR}{dt} = -R + V_m + 10(B - A). \tag{21}$$

The three terms on the right hand side of (21) are passive decay, an excitatory volitional term, V_m, and a term, $B - A$, that may be either excitatory or inhibitory, depending upon task dynamics. In particular, the rehearsal signal, R, receives excitatory input from the volitional signal, V_m, indicating that the system is in the recall phase of the task. The rehearsal signal also receives excitatory input from slowly integrating velocity cells, B, and inhibitory inputs from quickly integrating velocity cells, A. During the initial phase of a movement, the fast cells, A, are more active as they closely track the bell-shaped velocity performance curve, causing inhibition by ($B - A$) of the rehearsal signal. As the movement nears completion, the fast cells, A, closely track the decline in velocity at the end of the bell-shaped velocity curve while the slow cells, B, still encode the peak velocity, so the net effect on ($B - A$) is excitatory. A net positive rehearsal signal (R) opens the rehearsal gate in equation (14), thereby allowing inputs from the motor plan field (F) to update the plan selection field (S), and initiates selection of the next movement before current movement completion (i.e., achieves anticipatory movement selection).