ENCYCLOPEDIA OF COGNITIVE SCIENCE

2000

Computational models of working memory

working memory#computational models#serial order

Lewandowsky, Stephan

Stephan Lewandowsky

University of Western Australia, Nedlands, WA, Australia

Farrell, Simon

Simon Farrell

University of Western Australia, Nedlands, WA, Australia

Comparison of localist, distributed, and rule-based models of working memory gives insight into essential processes underlying the short-term retention of information.

Varieties of Memory

Memory researchers often distinguish between two manifestations of memory, one dedicated to the retention of information for very short periods, known as Short-Term Memory (STM) or *working memory*, and one dedicated to long-term storage, known as long-term memory (LTM). Working memory is involved when information is needed briefly but immediately; for example, when dialling a telephone number after looking it up in the directory. Long-term memory, by contrast, is involved when information needs to be available for repeated access across larger timescales, for example one's PIN.

Many theorists view working memory as a distinct functional and structural entity that requires examination and explanation in its own right. In support, many empirical dissociations between working memory and LTM have been reported. For example, whereas working memory is adversely affected by phonological similarity (e.g., memory for the list B D T V G is poorer than for the list T K X S M), this variable has little effect on LTM. Conversely, semantic similarity has little effect in working memory whereas its effects on LTM can be pronounced.

Accordingly, we restrict our focus to models of working memory, in particular those that are formulated at a *computational* rather than verbal level. Computational models provide a quantitative account of the data (e.g., through computer simulation) and thus offer more theoretical precision than verbal models. Although contemporary

computational models encompass a diverse range of theoretical approaches and assumptions, they all focus on performance in the immediate serial recall task, in which individuals must repeat back short sequences of verbal items in the correct order.

Much of the research involving this task, in turn, was stimulated by Baddeley's (1986) pioneering verbal theory of working memory, in particular his conception of the so-called phonological loop. Many contemporary models retain at least an empirical connection with Baddeley's proposal.

Baddeley's Phonological Loop

In Baddeley's (1986) model of working memory, verbal information such as lists of digits, letters, or words is maintained in a system known as the *phonological loop.* The phonological loop involves two components: a *phonological store,* which is a capacity-limited receptacle of information that is subject to constant decay, and an *articulatory control process*, a rehearsal mechanism that counteracts decay by periodically refreshing the items in the phonological store. Because the representational code is phonological, similar-sounding items interfere with each other, thus producing the observed disadvantage for phonologically similar lists.

At a verbal level of explanation, these simple assumptions account for numerous results, for example the fact that memory span (the number of items for which immediate serial recall is possible) is a linear function of pronunciation time for the material. Memory span is smaller for items that take longer to pronounce, presumably because this increases their exposure to decay in between rehearsals. The phonological loop also accommodates the observed elimination of this effect when subjects recite irrelevant material during list presentation. Under these "articulatory suppression" conditions, memory span for visually presented lists is poor but equal for items with short and long pronunciation durations, presumably because the rehearsal that could prevent trace decay is suppressed.

Working Memory As Active Localist Units

Several computational models have been proposed that formalize some of the concepts arising from verbal theorizing within the phonological loop framework. Although these models handle a largely overlapping set of phenomena, they differ in the processes thought to underlie maintenance and retrieval of information. We first consider "localist" models, in which each item is represented by a distinct unit whose activation represents memorial strength.

A Localist Connectionist Network Model of The Phonological Loop

Burgess and Hitch (1999) presented a network model of the phonological loop that combines localist representations with a time-driven "competitive queuing" mechanism for ordered retrieval. Thus, when an item is presented for study, phoneme units that correspond to its pronunciation pattern are activated, and that activation is fed forward to a layer of representational units. The representational unit whose activation is greatest, because it is excited by all presented phonemes, is associated

with a temporal "sliding window" representation of context. Context is represented by numerous units that are organized in temporal sequence, such that a contiguous cluster is activated at any given time. Across time, the set of activated units slides along the temporal dimension, thus producing some overlap between the set of units that are active at proximal time steps. At each time step, the weights between the activated representational unit and the context units are strengthened through Hebbian learning. Connection weights are also assumed to decay with time, thus capturing the transient nature of the phonological loop.

At recall, the context signal is reset, and its subsequent advance along the original temporal sequence is used to cue retrieval. The context signal activates a given representational unit to the extent that it overlaps with the time signal present at study. Because temporally adjacent context representations overlap, several units will be partially activated. The resultant competition among representational units is resolved through a winner-takes-all process based on mutual inhibition. Once an item has been selected for report, its associated output phoneme pattern is activated, which simulates spoken recall. Immediately upon recall, the representational unit is suppressed and thus removed from subsequent competitions. This item selection process is known as competitive queuing.

Several components of the model can be linked to Baddeley's verbal theory: Competitive queuing implements (part of) the articulatory control process within the phonological loop, and the phoneme units that activate representational units instantiate the phonological store.

The model handles the effects of pronunciation time and articulatory suppression that were cited earlier in connection with Baddeley's model. Moreover, in crucial distinction to other computational models, Burgess and Hitch can accommodate a large body of neuropsychological literature by simulating various different types of brain damage through selective lesioning of connections. Finally, the model handles many basic aspects of serial recall, such as the bowed shape of the serial position curve, error patterns, and the like. An earlier version of the model, using the same basic architecture and representation of context, also accommodated the effect of temporal grouping of items on the serial position curve.

Two drawbacks of the model can be identified on the basis of its complexity and its mixing of representational approaches. The model mixes representational approaches because representational units are localist—each item is represented by a single unit—whereas context is represented in a distributed manner. It is unclear on what a priori grounds this mixed representation could be justified. The model is complex because it includes four sets of weights that, respectively, connect context to items, input phonemes to items, items to output phonemes, and input to output phonemes. All but the last set of connections are adjustable through Hebbian learning, and each adjustable set in turn involves a plastic short-term weight that decays over time, plus a static long-term weight. A reasonably large number of parameters is required to govern and coordinate these multiple sets of weights, which renders it difficult to assess the overall parsimony of the approach and the power of its core assumptions. The latter issue can be examined by considering a much simpler model that also uses localist representations and competitive queuing.

The Primacy Model: Parsimonious Competitive Queuing

The Primacy Model (Page and Norris, 1998) implements competitive queuing without a context or timing signal. At study, localist item nodes are activated to an extent determined by the position of the corresponding item in the list. The first list item is maximally activated, and the activation of all other nodes decreases across successive serial positions, thus setting up a *primacy gradient*. Paralleling the presumed properties of the phonological loop, the activation of a node begins to decay immediately after the item has been presented. If rehearsal occurs, decay is reversed and activations are restored to their original levels as dictated by the primacy gradient.

At recall, the strongest item is reported and then immediately suppressed. Thus, because the first list item is maximally active, it will also be recalled first. Because recall is followed by suppression, that item is unlikely to be recalled again, and the second most active item is recalled next. This process of recall followed by suppression continues until all list items have been recalled. To allow the model to generate errors, noise is introduced into activations before selecting an item for report.

The model, though relying on very few parameters, can account for many results: It predicts the standard serial position curve, with an advantage for items at the start of the list (the primacy effect) and a similar, though smaller, advantage for the last few items in a list (the recency effect). The model also predicts the associated patterns of transposition errors, in particular the "locality constraint," which refers to the fact that items tend to be reported in proximity to their correct position. The model also handles modality effects (recall is enhanced for words presented auditorily), the effects of pronunciation time reviewed earlier, and list length effects (recall is worse for longer lists).

One shortcoming of the Primacy Model concerns "fill-in." Fill-in refers to the observation that if an item is recalled too early (e.g., Item 2 is recalled first), the displaced item is more likely to be recalled next than the item following the transposed item (i.e., Item 1 is more likely to be recalled next than Item 3). The Primacy Model handles the phenomenon of fill-in, but greatly over-predicts its extent, as an earlier item not recalled will be a formidable competitor at the next output position.

Another limitation of the Primacy Model is that it cannot account for grouping effects. Temporal separation of a list into distinct small groups is known to change the serial position curve and patterns of transpositions (see below); this is often taken as evidence of a hierarchical or multi-level representation. As the Primacy Model can only order items along the single dimension of node activations, grouping effects are beyond its purview. A model that overcomes these limitations is the Start-End Model (SEM).

SEM: The Start End Model

SEM (Henson, 1998) assumes that at presentation, a new token is created in short-term memory that contains a representation of the list item and information about its list position. That positional context information is provided by a start marker and an end marker whose strengths decrease and increase, respectively, across list positions.

When considered together, the strengths of the start and end marker define a unique list position, and those strengths are stored together with the list item. At retrieval, the positional context for each position is reinstated, and the token whose encoded context is most similar to the cue is recalled. Thus the success of recall depends on the overlap between the positional context provided by the start and end markers for that position and all other possible positions. As in the Primacy Model, the assessment of overlap is noisy, and report of an item is followed by its suppression.

SEM matches the explanatory power of the Primacy Model and additionally handles retention interval effects and the finding that protrusion errors (i.e., incorrectly recalling an item from list *n*-1 on list *n*) tend to maintain their position from list *n*-1. Notably, SEM gives a thorough account of grouping effects. Grouping increases overall list recall and engenders a scalloped serial position curve, with small primacy and recency effects within each group. SEM accounts for these grouping effects by incorporating two sets of markers, item markers and group markers. SEM also handles the intricate pattern of transposition errors in grouped lists, for example the fact that terminal group items are likely to transpose with each other even if group sizes are unequal. For example, in a list presented as 3-4, the third item in the first group, when erroneously recalled in the second group, is more likely to be reported in group position 4 than position 3. This supports SEM's assumption of relative coding involving end markers, and it simultaneously casts doubt on absolute coding, such as in the Burgess and Hitch model, which would predict that between-group transpositions should maintain their position within the group.

Most problematic in SEM is the end marker, whose strength increases exponentially across serial positions, with a predetermined maximum value for the last input position. This requires that list length be known ahead of presentation, which is clearly an unreasonable assumption as the data remain largely unchanged even if people cannot anticipate list lengths. Henson and Burgess (1997) suggest a solution to this problem that involves competition among multiple autonomous oscillators of differing frequency. When list presentation is complete, only those oscillators retain a role in item coding whose distinctiveness has reached its peak at the end of the list.

Working Memory As a Distributed Network

Distributed models differ from the preceding theories by assuming that items are represented not by identifiable single units but as patterns of activation across a collection of units. Typically, each unit is involved in the representation of several items, and information storage is accomplished by adjusting the weights between layers of such units.

Unlike some of the localist models, distributed models typically reject any explicit theoretical link with the phonological loop and its notions of decay and rehearsal. Instead, distributed models typically ascribe forgetting to interference. The models also postulate that memory is inherently associative and thus requires some type of cue to elicit recall.

OSCAR: Oscillator Based Model of Associative Recall

In the OSCillator-based model of Associative Recall (OSCAR) of Brown et al (2000), study items are represented by vectors that consist of many elements. Each element takes on an activation value that is randomly sampled from a Gaussian distribution. Upon presentation, each item is associated to a contextual cue and, through Hebbian learning, is superimposed upon all earlier memories in a matrix of weights. The contextual cue is provided by a collection of temporal oscillators with differing frequencies whose overall activation pattern uniquely specifies a given point in time.

At recall, the temporal signal is reproduced by first resetting the oscillators and then letting them re-evolve. Items are retrieved by cueing the memory matrix with the temporal context at each time step. Owing to the distributed representations and the use of Hebbian learning, the retrieved item is "noisy" and needs to be "redintegrated" into an overt response. (Redintegration is defined as the process by which partial memorial information is disambiguated and hence rendered available for output). In OSCAR, redintegration is performed by comparing the retrieved item to a pool of possible responses and choosing the one that is most similar. Once a response has been chosen, it is typically suppressed and thus unavailable for further report.

The explicit representation of temporal context is related to the approach chosen by Burgess and Hitch (1999), and it allows OSCAR to handle time effects like recency judgments and various retention intervals. However, the strict adherence to temporal oscillators brings the model in conflict with the earlier data that between-group transpositions tend to maintain their relative position within a group. Another limitation, typical of distributed models, is the fact that redintegration and response suppression affect a different form of representation (i.e., discrete responses among the pool of competitors) than that used to encode serial order (i.e., the weight matrix). Below, we present a recent solution to this problem.

TODAM: Theory of Distributed Associative Memory

Brief consideration must be given to the Theory Of Distributed Associative Memory (TODAM; Lewandowsky & Murdock, 1989). Although the theory has been clearly superseded by contemporary approaches, the reasons for its obsolescence are informative. In contrast to all models discussed so far, TODAM rejects the idea of positional encoding; instead, successive items are associated with each other in a chain. Recall involves probing memory with each successive list item. In its simplest version, this "chaining" view is immediately ruled out because it postulates that recall ceases with the first error, and thus cannot accommodate recency. TODAM circumvented this problem because its distributed representations, as in OSCAR, lead to "noisy" retrievals that require subsequent redintegration. This can generate recency because the retrieved approximation can correctly cue the next item even though redintegration yields an incorrect response. However, like OSCAR, TODAM implements redintegration by postulating that the retrieved response candidate is compared to discrete representations of all possible responses. As with OSCAR, this solution is unsatisfactory because the representational assumptions of an important retrieval component are at odds with the core properties of a distributed memory system.

TODAM handles many of the serial order effects reviewed in the context of the earlier models. Notwithstanding that apparent success, several recent studies cast doubt on the plausibility of chaining as a mechanism for ordering items (e.g., Henson et al, 1996). Chaining of items predicts catastrophic errors in lists of mixed confusable and non-confusable items, as the large overlap between similar items should cause massive interference among cues, and recall of the following items should be impaired. The data very clearly do not exhibit this pattern, and the failure of TODAM thus strongly suggests that inter-item associations do not play a crucial role in the representation of serial order information.

Redintegration with Distributed Representations

The redintegration difficulties cited in connection with OSCAR and TODAM have been, at least in principle, resolved by a recent dynamic attractor model (Lewandowsky, 1999). This model was not intended as a complete description of serial recall, but instead provides a mechanism for the redintegration of the noisy outputs of distributed models, such as TODAM and OSCAR, into overt responses. The model redintegrates a partial response by iteratively feeding it back into a weight matrix composed of the self-associations of all study items. Across iterations, the partial response is mathematically guaranteed to be disambiguated into *some* overt response, though the probability with which the correct response is chosen depends (roughly) on the similarity between the partial response and the correct item.

It is noteworthy that Lewandowsky's (1999) redintegration model, with minimal assumptions about how partial information is retrieved from memory, could by itself account for many aspects of serial recall. The model could produce the recency portion of the serial position curve, the correct shape of the transposition gradient, and the relative incidence of other errors across list and output position.

Working Memory As Rule-Based Models

In contrast to all foregoing models, rule-based models represent a list as a collection of propositions, and as such are only tenuously related to Baddeley's (1986) notion of the phonological loop. The primary model of this type is the ACT-R theory of Anderson (Anderson & Matessa, 1997) according to which a list is encoded in a hierarchical structure composed of propositions that encode the identity of items and item-in-group and group-in-list position.

Retrieval is coordinated by "production rules", which are condition-action pairs that execute particular functions when their conditions are true. Errors occur through partial matching of the condition of a production; hence items sharing similar codes will be co-activated, thus leading to potential confusion when noise is incorporated.

ACT-R handles the standard serial position curve, list length effects, the effects of pronunciation duration, the pattern of transpositions, phonological similarity effects, and the role of articulatory suppression. Some of these effects (e.g., list length) occur because there is a limited "pool" of activation, such that the activation available per item decreases as list length increases. Importantly, the model also accommodates the scarce available latency data from serial recall tasks.

The strength of ACT-R is that it integrates explanations for serial recall performance in a wider theory of cognition. There are, however, also noticeable shortcomings. Although the model qualitatively explains many phenomena, an associated successful quantitative fit is lacking (see Figure 8 of Anderson & Matessa, 1997). In particular, ACT-R fails to capture the extent of primacy and recency in the standard serial position curve.

Summary

Several overarching comments can be made about the state of computational modelling in working memory. First, there is little connection between the introspectively appealing notion of rehearsal that was central to Baddeley's (1986) phonological loop and the mechanisms embodied in current models. Even the theory that is most closely allied to the phonological loop, the model by Burgess and Hitch (1999), does not contain a process that retains the intuitive character of sub-vocal rehearsal. Our preferred interpretation of this state of affairs is that introspection and intuition can be poor guides to theorizing.

Second, there are some interesting overriding similarities between theories. For example, virtually all models assume some form of response suppression to prevent the same unit of information being repeatedly accessed. Another unifying assumption in many of the models is the superiority of encoding of early list items, which results in the primacy effect. In OSCAR, TODAM, SEM, the Primacy Model, and Lewandowsky's (1999) redintegration model, this is explained using some parameter which causes the quality of storage to decay across serial positions. A similar argument applies to the phonological similarity effect, which most models explain by relying on several stages to allow independent item and order confusions. This strategy was followed by the Primacy model, SEM, the Burgess and Hitch (1999) model and ACT-R.

The differences between the models are also enlightening. For example, it appears highly unlikely that a model without any kind of associations (e.g., the Primacy Model) will be able to account for grouping effects. Similarly, a comparison of TODAM to the other models suggests that pure chaining cannot underlie serial recall and that positional coding is the more likely alternative. A comparison of the Burgess and Hitch model and SEM furthermore suggests that that positional encoding is relative not absolute.

At the outset, we mentioned the prevalence of the distinction between short-term (or working) memory and long-term memory. Some of the present models characterize STM as a separate system, differing from LTM in processes and representation of items (e.g., Primacy Model, SEM). In consequence, these models probably cannot accommodate the ample evidence that LTM can affect performance on working memory tasks (e.g, Hulme et al., 1997). The issue of an LTM contribution is of importance because it relates to the purpose of working memory. There is now much empirical evidence that a separate verbal short-term system plays an integral role in vocabulary acquisition, or that vocabulary acquisition and immediate serial recall have a common substrate (Baddeley et al, 1998). Although the models discussed here have not been applied to vocabulary acquisition, improved knowledge of how temporal

order is represented in the serial recall paradigm may be critical to our understanding of language development.

References

Anderson JR and Matessa M (1997) A production system theory of serial memory. *Psychological Review* **104**: 728-748.

Baddeley AD (1986) *Working memory*. Oxford: Oxford University Press.

Baddeley AD, Gathercole SE and Papagno C (1998) The phonological loop as a language learning device. *Psychological Review* **105**: 158-173.

Brown GDA, Preece T and Hulme C (2000) Oscillator-based memory for serial order. *Psychological Review* **107**: 127-181.

Burgess N and Hitch GJ (1999) Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review* **106**: 551-581.

Henson RNA (1998) Short-term memory for serial order: The Start-End Model. *Cognitive Psychology* **36**: 73-137.

Henson RNA and Burgess N (1997) Representations of serial order. In: Bullinaria JA, Glasspool DW and Houghton G (eds) *4th Neural Computation and Psychology Workshop*, pp. 283-300. London: Springer.

Henson RNA, Norris D, Page MPA and Baddeley D (1996) Unchained memory: Error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology* **49A**: 80-115.

Hulme C, Roodenrys S, Schweickert R, Brown GDA, Martin S, and 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, and Cognition*, **23:** 1217-1232.

Lewandowsky S (1999) Redintegration and response suppression in serial recall: A dynamic network model. *International Journal of Psychology* **34**: 434-446.

Lewandowsky S and Murdock BB (1989) Memory for serial order. *Psychological Review* **96**: 25-57.

Page MPA and Norris D (1998) The primacy model: A new model of immediate serial recall. *Psychological Review* **105**: 761-781.

Further Reading

Conway MA (ed) (1997) *Cognitive models of memory.* Cambridge, MA: MIT Press.

Gathercole SE (ed) (1996) *Models of short-term memory*. Hove: Psychology Press.

Gupta P and MacWhinney B (1997) Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language* **59**: 267-333.

Levy JP, Bairaktaris D, Bullinaria JA and Cairns P (eds) (1995) *Connectionist models of memory and language*. London: UCL Press.

Miyake A and Shah P (1999) *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge: Cambridge University Press.