Proc 23rd Annual Conference Cog Sci Soc: 738- 743, 2001

Remembering to forget: Modeling inhibitory and competitive mechanisms in human memory.

Mike W Oram (mwo@st-andrews.ac.uk) and Malcolm D. MacLeod (mdm@st-andrews.ac.uk)
School of Psychology, University of St. Andrews, St. Andrews
Fife, KY16 9JU, UK

Abstract

Given the importance attached to memory in everyday life, the inability to recall items on demand can be problematic. An apparently ironic phenomenon has been identified, however, which suggests that in addition to retrieving desired memories, the act of remembering inhibits or suppresses related memories. We show here that a competitive model, designed to investigate the development of the cortical visual system, provides an explanation for the suppression of some memories as a consequence of remembering others. We confirm a number of specific predictions based on our model as to when retrieval-induced forgetting effects should or should not occur. The model suggests that the mechanisms by which memories are formed and adapted may also underlie retrieval-induced forgetting effects. In addition to having important practical implications, the model provides a theoretical base for the transfer of theories and ideas between two separate levels (cortical processing and memory formation and adaptation) of understanding brain function.

Introduction

Recent evidence suggests that far from being a detrimental process, forgetting has an adaptive role (Anderson & McCulloch 1999; Bjork 1989; Macrae & MacLeod 1999). When trying to remember a specific memory, available retrieval cues are often insufficiently specified to the extent that related but unwanted information is also accessed. This unwanted information can interfere with our ability to retrieve the information we wish to recall. A potential solution to this problem is through the temporary suppression or inhibition of related material (Anderson & McCulloch 1999; Anderson & Spellman 1995; Anderson et al 1994; Anderson, Bjork & Bjork 2000; Bjork et al 1998; Ciranni & Shimamura 1999; MacLeod & Macrae 2001; Macrae & MacLeod 1999). Importantly, this temporary suppression of related memories – retrieval-induced forgetting – occurs without the need for explicit cues to forget and can therefore be considered an intrinsic part of the act of remembering (Anderson & Spellman 1995; Anderson et al 1994; Macrae & MacLeod 1999). Other explanations, such as output interference (where items recalled early in a list can interfere with the retrieval of subsequent items) have been eliminated as potential explanations for this phenomenon using a variety of methods. Direct evaluation using statistical techniques have shown that there is no tendency for the retrieval-induced forgetting effect to be larger for those participants who recalled practised items first (MacLeod in press; MacLeod & Macrae 2001; Macrae & MacLeod 1999). More direct evidence that an inhibitory process is involved comes from the demonstration that temporary suppression is observed in all items that are related (whether by initial set or other semantic links) to the suppressed items (i.e. second order inhibition, Anderson & Spellman 1995).

Retrieval-induced Forgetting

In an experiment showing retrieval-induced forgetting, participants are typically given two sets of information to remember regarding two separate categories (*A_1, A_2,…,A_10, B_i, B_ii, B_iii,…,B_x, e.g, 'John_cheerful, John_tolerant,…; Bill_vigorous, Bill_sensible,…)* followed by a retrieval practice session on a subset of items from one of the lists (the retrieval practice or RP set, *A_1, A_2,…A_5*, e.g. complete the following: *'John_ch____')*. Following a distracter task (name as many capital cities as you can), participants are asked to recall as many of the items as possible.

Figure 1 shows the pattern of results from such an experiment (see Methods). A greater proportion of the practised items (RP+, left bar) were recalled than unpractised items in either the same set (RP-, middle bar) or in the unpractised set (NRP, right bar). This enhancement (RP+ versus NRP) shows the facilitatory effect of practice on subsequent recall. Retrieval-induced forgetting is evidenced by the fact that recall performance of the non-practised items in the practised set (RP-) was worse than the recall of non-practised items in the non-practised set (NRP). Thus, retrieval-induced forgetting is a selective suppression of related items and not a general suppression of all memories (Anderson & McCulloch 1999; Anderson & Spellman 1995; Anderson et al 1994, 2001; Bjork et al 1998; Ciranni & Shimamura 1999; MacLeod & Macrae 2001; Macrae & MacLeod 1999). An output interference explanation would predict that retrieval-induced forgetting effects would be higher where there was a tendency to recall RP+ items early in the list. As noted in the introduction, retrieval-induced forgetting is not due to items from the practiced subset (RP+ items)

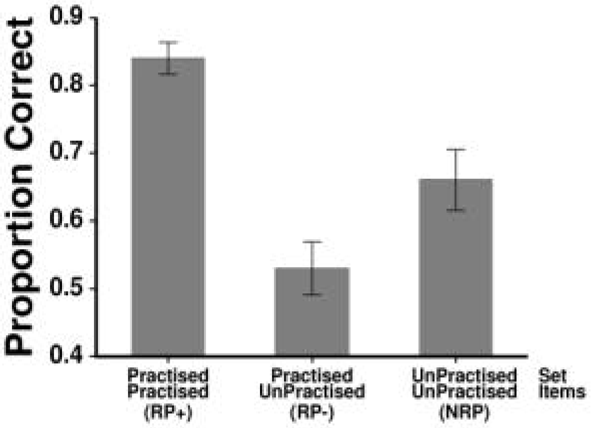

Figure 1: Retrieval-induced forgetting. Mean (±SEM, n=20) of the proportion of items remembered in each of each item type (RP+, RP- or NRP). The recall of unpractised items in the practised set (RP-) was less ($p < 0.05$) than the recall of the unpractised items in the unpractised set (NRP). Overall effect of conditions $F_{[2,38]}=28.3$, $p < 0.0005$.

being recalled first during the free recall task. (Anderson & Spellman 1995; MacLeod in press; MacLeod & Macrae 2001; Macrae & MacLeod 1999).

As the act of remembering during practice can selectively suppress memories for related but unpractised items, retrieval-induced forgetting must be influenced by the relationships between the items established during memory formation (Anderson & McCulloch 1999; Anderson & Spellman 1995). We were therefore interested in whether or not the mechanisms underlying retrieval-induced forgetting were different from the mechanisms that established the memories. We begin by considering, in broad terms, the required properties of a model consistent with the experimental data on retrieval-induced forgetting. As retrieval-induced forgetting is undirected (Anderson & McCulloch 1999; Anderson & Spellman 1995; Anderson et al 1994; Macrae & MacLeod 1999), learning should be unsupervised. Also, as retrieval-induced forgetting occurs with both semantic (Macrae & MacLeod 1999) and episodic memories (Ciranni & Shimamura 1999), the model should show unsupervised learning of both semantic and episodic-like memories. Finally, as inhibitory mechanisms are implicated, the model should contain inhibitory or competitive processes. We first show that a model consistent with this broad outline shows retrieval-induced forgetting. We then use the model to formulate three predictions of when retrieval-induced forgetting will not be observed. These predictions are verified experimentally.

Methods

We employed a computational approach to aid understanding of the role of inhibitory mechanisms in mental life. Computational testing of psychological theories can provide a powerful conceptual framework from which principled sets of research questions can be derived. However, using computational models in this way is not straightforward. The high number of degrees of freedom can lead to over-fitting the data and hence offer neither explanatory power nor generalisation to other scenarios. Hence, the observation that a model can fit experimental data is insufficient to validate the underlying processes within the model. In addition, results from a model developed around underlying psychological processes will be restricted in interpretation to the assumed underlying psychological processes: such a model can determine whether the assumed processes could underlie observed phenomena, but is weak at determining whether the assumed processes are actually in operation and important.

We address the caveats of using computational models to investigate psychological processing in two ways. First, selection of the category of model is made in broad terms without specific implementation to match observed psychological phenomena. If such a model is observed to produce the phenomena of interest, predictions from changing parameters in the model can then be validated with experimental data. The experimental validation of predictions overcomes, at least partially, the difficulties associated with many degrees of freedom that, in turn, gives rise to over-fitting the experimental data. Second, we assume that if the model reflects the psychological processes in a meaningful way, the parameters of the model will relate to psychological processes. This is not simply that the output of the model relates to the phenomena of interest, but that the parameters relate to underlying psychological processes. If the parameters of a model can be related to psychological processes, then the model may provide insight into how these processes interact.

Simulation methods

Damage to cortical tissue appears necessary for retrograde amnesia, implying that the neural representation in cortex correlates with the long-term memory. As inhibitory mechanisms are implicated in both the formation and functioning of neural representation (Oram and Perrett, 1994; Desimone and Duncan, 1995) and cognitive interactions between those representations (Anderson et al., 1994; Anderson and Spellman, 1995; Ciranni and Shimamura, 1999), we chose to investigate whether inhibitory processes involved in the formation of representations/memories could also underlie the interactions between representations/memories revealed by retrieval-induced forgetting.

The model consisted of two sets of 10 input nodes representing the individual items and two input nodes representing the set identifiers. The 22 input nodes were fully connected to the 10 output memory nodes, initially with random weights (0..1). Each node had an associated trace activity, Tr, which was dependent on the node's Tr at the previous time step and the node's current activity, Act: $Tr_{(time+1)} = (1-\delta)Tr_{time}+\delta Act$. The trace activity time constant δ was set at 0.5, with similar results obtained for δ=0.2 to δ=0.8. Weights between input i and memory node j were set randomly (0…1) with updating (learning) based on the trace activity, $\Delta Wt_{[i,j]} = \alpha(Act_i - Wt_{[i,j]})Trace_j$. The weight change rate, α, was 0.01 (similar results were obtained for α=0.001 to α=0.2). The ($Act_i - Wt_{[i,j]}$) ensures that the weights are bounded (-1..1). Initial training consisted of setting the activities of the input node corresponding to one of the input items to 1, calculating the activity of the memory nodes, updating the weights, then resetting the activity of the input node to 0, then "presenting" another input item. The activity of the set node associated with each input item was set to 1 while items within the set were presented. Retrieval practice was run in an analogous way for one half of the items in set 1, except that activity of the item nodes was set at 0.5 representing the partial cueing in the experimental paradigm. The representational strength was calculated as the activity in the item nodes following activation of 1.0 of the set node. Weight change was calculated as the change in the representational strength from after training to after retrieval practice. The change was normalised by dividing by the representational strength after training.

Experimental methods

Following Anderson et al (1994), the study comprised four phases: study, practice, distracter and final test. Participants were presented with ten items of information presented individually for 5s about two witness statements (one concerning a personal theft and the other a bank robbery). The practice phase followed immediately after the study phase. Participants were presented with five questions about a subset of items concerning one of the witness statements (RP+ items). Each question was presented three times. Counterbalancing and randomisation of question order ensured that each of the items appeared equally often in the RP+, RP-, and NRP conditions. Participants were then given a 5-min distracter task (write down as many capital cities as you can). Finally, participants were given a surprise free recall task in which they were required to recall as much of the information contained in the two statements. The number of correctly recalled items was noted for each group (RP+, RP- and NRP) and converted to proportion correct by dividing by the number of items in each group (RP+=5, RP-=5, NRP=10 and Figure 1).

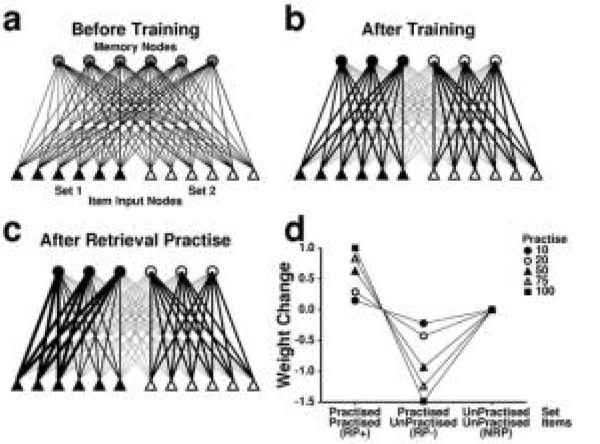

Figure 2: Modelling retrieval-induced forgetting. The weights between item nodes and memory nodes (representing the strength of the memory) are depicted by the thickness of the connecting line. The grey scale of the memory nodes indicates the strength of connection between the set node (not shown) and the memory nodes. A: Before training, the weights between input and memory nodes were random. B: Weights after the training. C: Retrieval practice re-distributed the weights linking the set 1 items with the memory nodes while leaving the weights on set 2 unaltered. D: Change in the weights representing the memories of the items and their set membership. As practice increases from 10 to 100 iterations, the strength of the memory inhibition also increases.

Results

We adapted a fully connected single layer unsupervised competitive model that forms both semantic and episodic like memories by learning from both past and present activity (Foldiak 1990, 1991; Oram & Foldiak 1996). The model consists of two sets of input items, each containing 10 items. Two additional inputs were used to indicate the training set. Initial weights from the input to output nodes were set randomly. Competitive interactions were modelled using a winner-take-all implementation (Foldiak 1991; Oram & Foldiak 1996) between the 10 nodes in the output layer. The network we adapted loads each output node with equal share of the input variance (Foldiak 1990; Oram & Foldiak 1996). The trace activity (Tr) imparts a structure to the inputs in the form of temporal co-variance between items. This co-variance results in the equally distributed input variance being parceled into equal variances associated with the different input sets and, within each set, an equal representation of the individual items. Thus, each output node learns part of the co-variation between a "set" node and the "item nodes". The resulting representation is best described as sparse, being neither fully distributed nor local. Sparse representations have the benefits of both distributed and local representations and seem to describe accurately cortical representations. The greater the number of output nodes, the sparser the representation. Qualitatively similar results are obtained when the number of output nodes varies from 4-30 output nodes.

There were two phases to training: in the 1st phase, the model was sequentially presented with each of the items with the items set membership also activated. This is analogous to the initial learning phase of retrieval-induced forgetting paradigms. In the 2nd phase, the model is sequentially presented with half the items from one set partially activated (the retrieval practice phase). The changes in the strength of the model's representations of items at different stages during simulated retrieval-induced forgetting are shown schematically in Figure 2a-d (thick lines indicate a strong link, thin lines indicate a weak link). Before training (Figure 2a) the weights are random and small. Learning rules based on recent as well as current activity, such as those employed here, learn temporal relationships between inputs (episodic-like memories) as well as relationships between nodes with concurrent activity (semantic-like memories). This allows the individual set-item relationships and the relationships between the different items within the same set to be learned. The inhibitory competition between nodes keeps the set-item representations of different sets of inputs separate (Foldiak 1991; Oram & Foldiak 1996). After training (Figure 2b) the representation of the

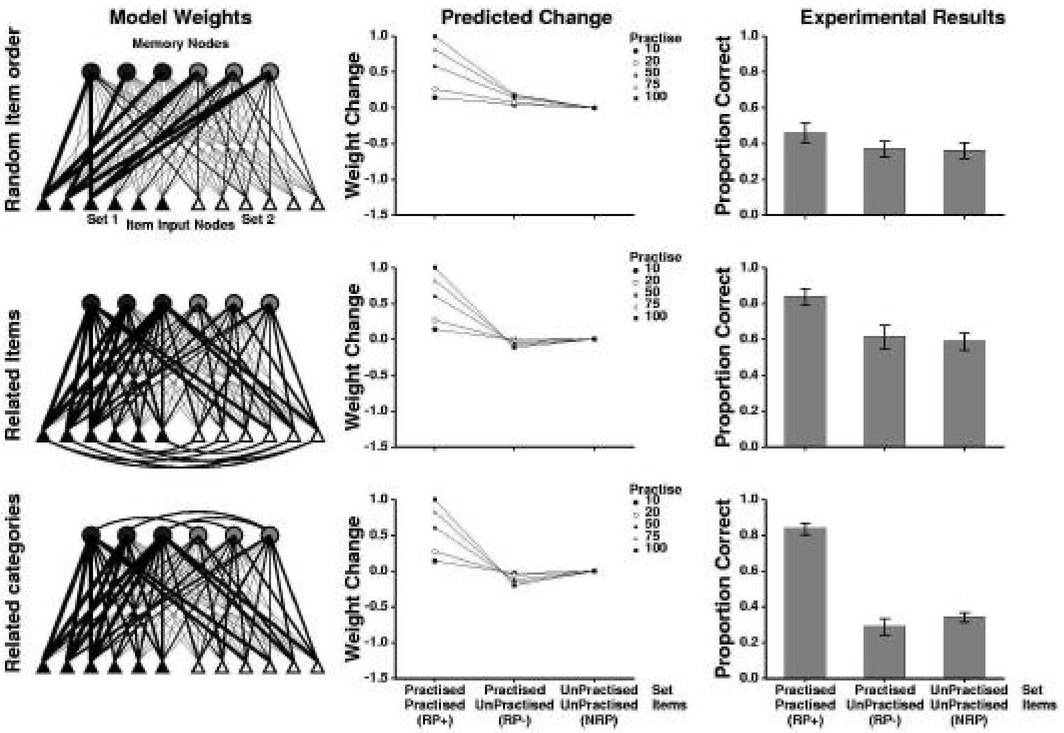

Figure 3: Predicting the disruption of retrieval-induced forgetting. The three rows show different conditions under which the model predicts that retrieval-induced forgetting will not occur and the results of experimental studies. The change in the representational strengths of the RP+, RP- and NRP items in the model following retrieval practice are shown in the middle column. The results of the experimental studies (mean proportion correct ±SEM in free recall for the RP+, RP- and NRP item types) are shown in the right column. Upper Row: Lack of coherence between items (random presentation of items). The model was run with trace activity time constant $\delta=1$ (otherwise experiment as in Figure 1) producing overlapping representations of set 1 and set 2 (left). Accuracy of predicted recall was reduced compared with items which were temporally coherent. The absence of retrieval-induced forgetting (compare top middle with Figure 2D)was confirmed experimentally ($p > 0.5$, right). Middle Row*:* Effect of semantic links between items. Direct connections (strength = 0.5) between input items in set 1 and set 2 produces overlapping representations of set 1 and set 2 (left). Retrieval-induced forgetting was so attenuated that it was predicted to be undetectable experimentally (middle). This was confirmed experimentally ($p > 0.2$). Lower Row: Effect of confounding by category. When connections between memory nodes (25% chance) were included (strength = 0.5), item-set memories showed overlap between sets (left), which again predicts greatly attenuated effects of retrieval-induced forgetting (middle). This was also confirmed experimentally ($p > 0.05$).

items in the memory nodes is divided into two sets, with little or no overlap. Following retrieval-practice (Figure 2c), the strength of the representation of practised items (RP+) is increased (without the simulated retrieval practice, the representation of the RP- and RP+ items is equivalent). The strength of the representation of unpractised RP- items is reduced because of repeated occurrences of high activity in the memory nodes with no activity in the input nodes representing the RP- items. As the retrieval-practice did not activate the memory nodes associated with the NRP items, the strength of representation items in the non-practised set is not influenced by retrieval-practice. Figure 2d shows that the network predicts the phenomena associated with retrieval-induced forgetting: the expected recall of RP+ items is enhanced compared to the recall of the NRP items and, as with retrieval-induced forgetting, the recall of the RP- items is lower than the recall of the NRP items. Thus, a competitive model can show retrieval-induced forgetting effects. The model does not provide a direct prediction of any effect of the order in which items will be recalled: we use the strength of representation as our metric. Note, however, that although the representational strength would suggest free recall beginning with RP+ items, the retrieval-induced forgetting effect in the model is not due to any form of output interference, only the strength of representation.

Given the variable success experimenters have had in producing retrieval-induced forgetting effects, we were particularly interested in examining the conditions under which retrieval-induced forgetting would not occur. We wished therefore to evaluate the model by changing those model parameters which suggest a strong influence on retrieval-induced forgetting and experimentally testing the predicted effects. We chose to manipulate those model parameters that have readily identifiable psychological counterparts. If the model parameter that represents the degree of continuity between the items (trace activity) is reduced, the items in each of the different sets to be remembered do not form a coherent pattern after the initial training, and the clarity of the relationships between items and their sets is reduced (Figure 3, top left). The overlap between the representations of items in different sets predicts reduced levels of recall performance compared with the situation where the continuity between items is easily established. Following simulated retrieval-practice, the strength of both the RP- and the NRP item representations is reduced while the strength of the RP+ representations is increased, i.e., an absence of retrieval-induced forgetting (Figure 3, top middle). When items used in the initial experiment (Figure 1) were presented in random order such that no coherence between the items was evident, retrieval-induced forgetting did not occur. In addition, absolute performance levels were reduced compared to when the same items were presented in a coherent fashion (compare Figure 3, top right and Figure 1).

Links between individual input items of the different sets can be thought of as exemplar-exemplar links based on semantic relationships between item inputs. Activation of one item will lead to (partial) activation of those related items in the second set. The concurrent activation leads to item representations that do not map perfectly with the input set (Figure 3, center left), so that retrieval practice reduces the strength of representation of both the RP- and NRP items whilst increasing the RP+ representation (Figure 3, center). Semantic relationships between input items were obtained experimentally by using appearance descriptors concerning two individuals (e.g. *Bill_Nike trainers*, *Bill_Slim build…*, *John_Adidas trainers*, *John_Medium build*) as the input items (*trainers*, *build* etc forming explicit links). As predicted, retrieval-induced forgetting did not occur (Figure 3, center right). Finally, links between the representations of the item groups (the memory nodes) models the existence of pre-existing groupings involving the items. This can be thought of as the existence of indirect or implicit semantic links (exemplar-category-exemplar). The overlap of pre-existing groupings of the items of the different sets leads to the representation of single items being associated with both sets (Figure 3, lower left). The effect of confounding relationships between the memory nodes was examined by asking participants to learn representations of employees in different companies that were confounded by gender. Again, the prediction from the model was met: retrieval-induced forgetting did not occur (Figure 3, lower right).

Discussion

The results of these studies highlight two important aspects of memory formation and maintenance. First, we have shown a mechanism by which practice and revision (consolidation) of selected memories can lead to suppression of related memories but leave unrelated and unpractised memories unaffected (Figure 1). While others have noted the restricted occurrence of retrieval-induced forgetting (Anderson & McCulloch 1999), our model allows specific predictions to be made about both performance levels and the strength of retrieval-induced forgetting effects. The four predictions about performance in a cognitive memory task (Figure 3, middle column) were all tested and verified experimentally (Figure 3, right column). This suggests competitive models with learning based on past as well as present activity can help predict how, why and when these types of memory interactions occur. Second, the model suggests that the effects of practice and revision of selected memories are due to the same processes by which memories are first established and hence need not be regarded as separate cognitive processes. Support for retrieval-induced forgetting as an intrinsic property of memory formation comes not simply from the demonstration that a model can produce retrieval-induced forgetting effects without explicitly coding the effect, but also that the same model predicts the absence of retrieval-induced forgetting effects (Figure 3).

In day-to-day function, retrieval-induced forgetting is important because it allows the updating or alteration of memory without interference of or disruption to other memories. For example, remembering where you parked your car today rather than where you had parked it yesterday should not interfere with your memory of the shopping you need to do. This type of selective adjustment of memories has practical implications: police interview techniques could be adjusted to minimise the potential loss of pertinent information from witnesses; teaching the establishment of conceptual links between aspects of the curriculum should be emphasised with revision of all the related material; students who revise only part of their course may well be placing themselves at a disadvantage because of the active suppression of related memories. If, as our model suggests, retrieval-induced forgetting effects are intrinsic to memory formation, then a simple way of reducing susceptibility to this kind of forgetting is to create many links during initial learning – perhaps the reason why the development of complex schemata provides resistance to such forgetting (Anderson & McCulloch 1999).

We have shown that a competitive model reveals a potential mechanism allowing prediction of experimental data concerning the cognitive processes of memory formation and adaptation. Our model shares similarities with that of Bauml (1997). However, our model suggest the suppression normally attributed to retrieval processes could itself be part of the mechanism by which memories interact and are updated. The choice of model provides not only a potential explanation of memory formation and adaptation but also demonstrates that a mechanism proposed to describe the selectivity of single cells within extra-striate visual cortex (Foldiak 1991; Oram & Foldiak 1996; Oram & Perrett 1996; Wallis & Rolls 1997) can operate at the much coarser scale associated with episodic and semantic memories and their interactions. The existence of a single model that operates at both fine (single cell) and coarse (episodic and semantic memory) scales is appealing because it provides a medium for the transfer of theories and ideas between two different levels of approach to brain function and their subsequent testing.

Acknowledgments

We acknowledge Dr. J. Gerry Quinn for helpful discussion and reading of draft manuscripts.

References

Anderson, M.C., Bjork, R.A. & Bjork, E.L (1994) Remembering can cause forgetting: Retrieval dynamics in long-term-memory. *J Expl Psychol-Learn Mem Cogn* 20, 1063-1087.

Anderson, M.C., Bjork, R.A. & Bjork, E.L. (2000) Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychon Bull Rev*, 522-530.

Anderson, M.C. & Neely, J.H. (1996) in *Memory: Handbook of perception and cognition* (eds Bjork, E.L. & Bjork, R.A.) 237-313 (Academic Press, New York.

Anderson, M.C. & McCulloch, K.C. (1999) Integration as a general boundary condition on retrieval- induced forgetting. *J Exp Psychol Learn Mem Cogn* 25, 608-629.

Anderson, M.C. & Spellman, B.A. (1995) On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychol Rev* 102, 68-100.

Bauml, K.H. (1997) The list-strength effect: Strength-dependent competition or suppression? *Psychonom Bull Rev* 4, 260-264

Bjork, R.A. (1989) in *Varieties of memory and consciousness: Essay in honor of Endel Tulving* (eds Roediger, H.L. & Craik, F.I.) (Erlbaum, Hillsdale, NJ).

Bjork, E.L., Bjork, R.A. & Anderson, M.C. (1998) Varieties of goal-directed forgetting. In (eds J.M. Golding & C.M. MacLeod) *Intentional forgetting.* Mahway: NJ Erlbaum.

Ciranni, M.A. & Shimamura, A.P. (1999) Retrieval-induced forgetting in episodic memory. *J Exp Psychol Learn Mem Cogn* 25, 1403-1414.

Foldiak, P. (1990) Forming sparse representations by local anti-Hebbian learning. *Biol Cybern* 64, 165-170.

Foldiak, P. (1991) Learning invariance from transformation sequences. *Neural Comput* 3, 194-200.

MacLeod, M.D. (in press) Retrieval-induced forgetting in eyewitness memory: Forgetting as a consequence of remembering. *Applied Cognitive Psychology* In press.

MacLeod, M.D. & Macrae, C.N. (2001) Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychol Sci* 12, 148-152.

Macrae, C.N. & MacLeod, M.D. (1999) On recollections lost: When practice makes imperfect. *J Pers Soc Psychol* 77, 463-473.

Oram, M.W. & Foldiak, P. (1996) Learning generalisation and localisation: Competition for stimulus type and receptive field. *Neurocomputing* 11, 297-321.

Oram, M.W. & Perrett, D.I. (1994) Modeling visual recognition from neurobiological contraints. *Neural Networks* 7, 945-972.

Wallis, G. & Rolls, E.T. (1997) Invariant face and object recognition in the visual system. *Prog Neurobiol* 51, 167-194.