An Analysis of the Working Memory Capacity Paradox

Eddy J. Davelaar (e.davelaar@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck College
London, WC1E 7HX United Kingdom

Abstract

In the literature on working memory (WM), a paradox exists according to which very similar memory tasks provide support for very different estimates of working memory capacity. The current paper analyses the conflicting estimates of a capacity of 4+/-1 with a capacity of 1. To this end a dynamic process model of short-term recognition is used to generate data to which exponential speed-accuracy trade-off functions are fitted. The results show that even though the process model has a capacity larger than one, the exponential SAT functions indicate a one-chunk hypothesis. Further nested modeling reveals, counter to the dominant belief, that retrieval rate is insensitive to differences in WM capacity. The resolution of the WM capacity paradox lies in the choice of dependent measure.

Keywords: working memory capacity; speed-accuracy tradeoff; memory retrieval; model comparison.

Introduction

The last ten years have seen increased efforts in elucidating various aspects of working memory. Currently, there are several theories of working memory (see the chapters in Miyake & Shah, 1999) giving different explanations of behavioural data. Although many similarities exist among the theories, there are also important differences. In this paper, I will address the paradox of different estimates of working memory capacity and contrast the view that working memory can hold about 4 +/- 1 chunks (Cowan, 2001) with the view that the focus of attention is limited to 1 chunk (McElree, 2006). The paradox lies in the fact that the behavioural paradigms that provided different estimates are very similar – presentation of a sequence of words – whereas the dependent measure differs. I will use an activation-based model of working memory that has been applied to the list presentation paradigm (Davelaar, et al., 2005, 2006) and assess whether the model can reconcile the different views. Stated differently, is it possible that the estimate of 4 +/- 1 is compatible with the estimate of 1, when the paradigm-specific feature, i.e., the dependent measure, is taken into account?

The starting point is the paper by Nelson Cowan (2001) in which he reviewed a wide literature on attention and memory and concluded that the capacity limit or the focus of attention is around four chunks. Such a limit was suggested previously in a review by Donald Broadbent (1975) based on similar analyses of the literature. Furthermore, computational analyses using models such as the Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1980) supported the estimate of around four (Raaijmakers, 1982).

The commentaries based on Cowan's target article included empirical arguments supporting the view that the focus of attention is limited to one chunk (McElree & Dosher, 2001). This particular empirical argument focuses on the speed of retrieval from working memory and is central to the current paper. McElree and Dosher (2001) based their argument on data obtained using the response-signal speed-accuracy tradeoff (SAT) procedure. In this procedure, participants are presented with a sequence of words and receive a test probe after the final item. The participant has to indicate whether the test probe is one of the items in the just-presented sequence. Instead of freely responding, the participant makes a response as soon as a signal (e.g., a beep) is given. The profile of retrieval can be mapped out by employing a wide range of response signal delays. With very short delays, the participant is unlikely to have processed the test probe and performance is at chance. With a longer delay, performance rises above chance and with very long delays, performance asymptotes. The function that is traced by this procedure is called the speed-accuracy tradeoff function and can be described by or fitted with Equation 1 that involves three parameters: the intercept (T_0), the rate (s), and the asymptote (d'_{asy}).

$$d' = d'_{asy}\,(1 - e^{-s(t-T_0)}) \qquad \text{for } t>T_0,\ 0 \text{ otherwise} \quad (1)$$

The argument favouring the one-chunk hypothesis is as follows. Assume that the representation can either be in or outside the focus of attention. When it is in the focus of attention it is more readily accessible and should therefore lead to a faster rate of retrieval. This is measured by the rate parameter of the SAT function. Empirical studies consistently show (e.g., McElree, 1996; McElree & Dosher, 1989; Wickelgren, Corbett & Dosher, 1980) that the SAT function for the very last item has a faster rate than the SAT functions of the other items. In addition, the retrieval speeds for all pre-final items are equal. This suggests that the very last item is in the focus of attention, while the other items are not and thus that the capacity is limited to one item – the very last presented (or the very last processed McElree, 1998) item.

Initially, one would comment that it is possible that the most recent item is consistently in working memory, whereas the pre-final items reside in working memory with a lower probability. Therefore the estimated retrieval speeds for those items is a mixture of the fast and slow speeds, where the slow speed correspond with retrieval of presented items that are displaced from working memory (Cowan, 2001). The implied assumption underlying this view is that the probability of residing in working memory is a constant factor. Two objections to this assumption can be articulated. First, if a fixed-capacity buffer is used to encode a *sequence* of words, the probability of being in the buffer is highest for the most recent item. Thus theoretically, there is recency gradient *within* the buffer. Second, empirical observations show a recency gradient over the last four items for accuracy and reaction times (e.g., McElree & Dosher, 1989; McKone, 1995; Ratcliff, 1978), suggesting that if these items are in the buffer, a recency gradient must exist within the buffer.

To appreciate the complexities of these findings, consider that the encoding phase in the paradigms used by Raaijmakers (1982) and McElree and Dosher (1989) is identical but that the test phase differs. In addition, whereas Raaijmakers (1982) and Cowan (2001) focused on memory accuracy, McElree and Dosher (2001) focused on retrieval rate, which they argue provides direct evidence for distinct representational states. It should be said that the asymptotic accuracy of the SAT functions show a typical recency gradient. Therefore the paradox might be recast as a difference in opinion about what constitutes a proper dependent measure. This might well be the critical factor that prevents resolution of this central feature of working memory. The proposed way forward is to use a computational model with a capacity larger than one and produce the SAT functions. This requires (1) a *process* model of recognition memory that (2) implements a dynamic buffer, and (3) is capable of producing retrieval dynamics that can produce SAT functions. Several process models of recognition memory exist (Gillund & Shiffrin, 1986; Hintzman, 1984; Hockley & Murdock, 1987; McClelland & Chappell, 1998; Norman & O'Reilly, 2003; Shiffrin & Steyvers, 1997), but only a subset have been applied to SAT functions (Diller, Nobel & Shiffrin, 2001). Instead of readjusting the models to also include a dynamic buffer, the research strategy followed here is to extend a dynamic buffer model (Davelaar, et al., 2005; Haarmann & Usher, 2001) with a matching process that allows for a yes/no-recognition decision. This involves combining the dynamic buffer model with Ratcliff's (1978) diffusion model.

Model Description

The dynamic buffer model is based on the view that the content of working memory is the active part of long-term memory. More precisely, representations in consolidated memory, such as semantic long-term memory, phonological long-term memory (Baddeley, Gathercole & Papagno, 1997), and other modalities in long-term memory, are activated through sensory information. This activation is short-lived and would decay to baseline activation if there was not an active process that counteracts this decay. This process of active maintenance is a function of working memory (Baddeley, 1996) and has been called primary memory (Norman, 1968). The consequence of this process is that more than one representation can be activated simultaneously, albeit at different levels of activation. Previous work has shown that this model, which has many points of contact with Cowan's embedded processes framework (1995, 2001), is able to capture several observations in list memory paradigms. The core aspect of the model is the differential Equation 2 that governs the change of activation for every representation in long-term memory per timestep,

$$\frac{dx_i}{dt} = -x_i + \alpha F(x_i) - \beta \sum_{j \neq i}^{N} F(x_j) + I_i + \Phi(0, \sigma) \qquad (2)$$

where x_i is the internal activation of representation *i*, F = 1/(1+x) is the output activation function, α captures the process of active maintenance. When $\alpha = 0$, the model reduces to system with a capacity of one and is indistinguishable from theoretical models that purport to assume that only one representation can be active at any one moment (Brown, Neath & Chater, 2007; Howard & Kahana, 2002)[1]. All representations compete with each other through the inhibition parameter, $\beta = 0.2$, which governs the maximum capacity. Each representation receives activation, $I_i = 0.33$, from sensory processing levels. The activation dynamics is supplemented with zero-mean Gaussian noise with standard deviation, $\sigma = 1.0$. Representations that are active above a fixed threshold $\theta = 0.2$ interact with other aspects of the cognitive system. This includes episodic memory encoding and probe matching.

The diffusion model as used by Ratcliff (1978) is in essence a dynamic signal detection model and includes the mean drift rate, ξ, which represents the amount of match between the probe and the memory item. From trial to trial the amount of match varies and this variability is captured by the standard deviation, η, of the drift rate. When applying the diffusion model to behavioural tasks, the effective drift rate for a given trial is drawn from a normal distribution with mean υ and standard deviation η. For each unit of time, zero-mean Gaussian noise with standard deviation 0.1 is added to the mean drift rate causing the total amount of evidence indicating a match or mismatch to drift towards a boundary. When a match boundary is reached, system responds with a yes-response. When a non-match boundary is reached, a no-response is emitted. The original diffusion model has many more parameters and has been applied to a wide range of reaction time paradigms. Relevant to the current discussion is that the diffusion model has been

[1] So-called single-store models include some form of relative strength calculation. When reimplementing those models in a connectionist form in order to allow direct comparison, these models require a stage where multiple representations are active to allow for the ratio-rule type of calculation. An extreme version of this is where only one representation is allowed to be active during encoding, while multiple representations are active during retrieval (Sederberg, et al., 2008).

applied to the response-signal speed-accuracy tradeoff procedure (McElree & Dosher, 1989; Ratcliff, 1978, 2006).

The diffusion model takes the value for the drift rate from the dynamic buffer model. Specifically, the drift rate on each trial is the above-threshold activation for that representation. To produce SAT functions, the following two situations need to be explicated. First, when the response-signal appears and the diffusion process has not reached any boundary, the response is based on whether the process is moving towards the yes- or no-boundary. This represents making decisions based on partial information (see for discussion, Ratcliff, 2006). Second, when a boundary has been reached before the response-signal, the corresponding decision will be given at the time of the response-signal. The resulting decision probabilities are converted into d' scores and the full SAT functions are fitted with two version of Equation 1. In version 1, all parameters are free to vary across conditions, yielding 18 free parameters. In version 2, the reduced model that is supported by the empirical literature is used. This model has a fixed T_0 for all conditions and two different rates, yielding 9 free parameters.

The process model as described above was applied to a sequence of six words. Each of six representations was activated sequentially for 1,000 iterations. Then one of the six positions was probed and a SAT function created for that serial position by using response-signals at 100, 200, 300, 400, 500, 750, 1,000, 1,500, 2,000, and 3,000 iterations. Each serial position was probed 1,000 times at each of the ten response-signal delays. The effective capacity of the model is easily assessed by counting the number of representations that are active above threshold at t = 6,000 iterations. In order to address the possibility that different parameters obtained from the exponential SAT function are sensitive to different working memory capacities, the simulations are repeated for $\alpha = 0$ (no buffer), $\alpha = 1.8$ (small capacity), and $\alpha = 2.0$ (large capacity).

Simulation Results

Figure 1 shows a noise-less simulation of a sequence (with $\alpha = 2.0$). At time = 6,000, the very last item is the most active and activation levels decrease with the temporal distance of presentation. Figure 2 shows the frequency distribution of the activations for each of the items in Figure 1 at t = 6,000 iterations. As can be seen, items that are still in the activation buffer at time of test show a step-like function, with the very last item being more active than all other active items, which in turn have similar activation levels. The reason for this is immediately apparent when taking a closer look at Equation 2. Assume that at time of test, the activation level does not change and is above threshold. The resulting $F(x_i)$ is governed by α and β, leading to convergence of the activations. Only the very last item still receives external input, leading to a higher activation.

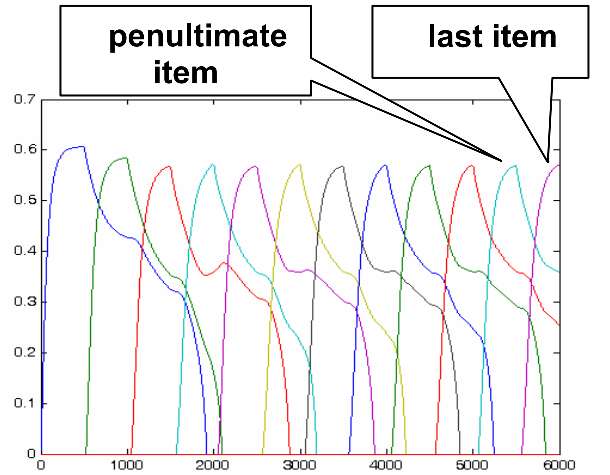

Figure 1. A noise-less simulation of 12 sequentially activated items. The x-axis indicates time in iterations. The y-axis indicates activation level, $F(x_i)$.

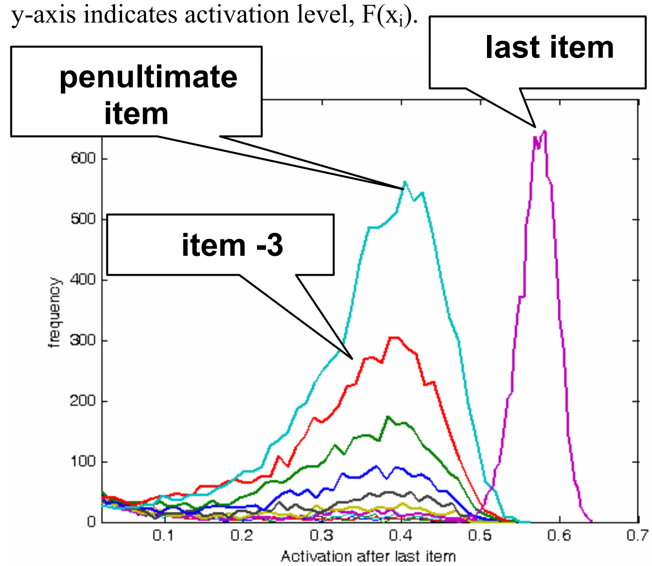

Figure 2. Frequency distributions of the activation levels of the 12 items in Figure 1 at t = 6,000 iterations.

The simulated data and corresponding best-fitting SAT functions for the simulation of $\alpha = 2.0$ are presented in Figure 3. Table 1 shows the parameter values of the best-fitting reduced model for each of the values of α. The models were fit by maximising the adjusted R^2.

Although the reduced model fits the data less well compared to the saturated model, the change in goodness of fit, ΔR^2, is negligible given the amount of variability present in real data. This supports the findings in the empirical literature that led to the one-chunk hypothesis. However, the model maintains multiple items at the time of test, as seen by the capacities. The capacity at $\alpha = 2.0$ is higher than at $\alpha = 1.8$.

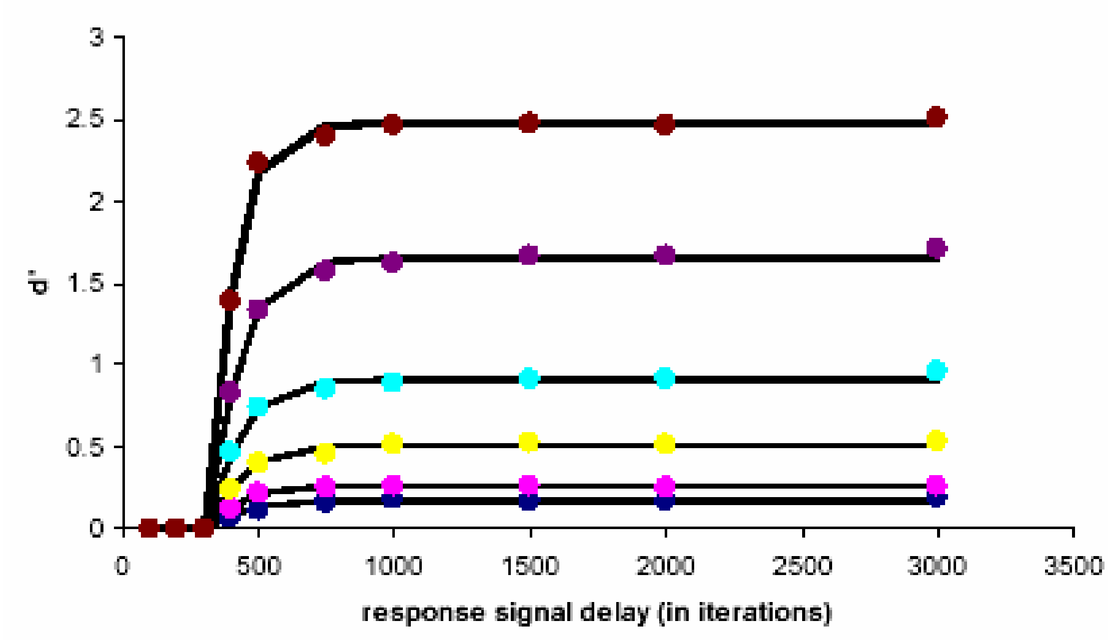

Figure 3. Simulation data and best-fitting reduced model for the simulation with α = 2.0.

Table 1: Parameter estimates for the 9-parameter exponential SAT function and the estimates of buffer capacity.

parameters	Serial position	simulation α = 0	α = 1.8	α = 2.0
d'_{asy}	1	0.015	0.014	0.173
d'_{asy}	2	0.028	0.031	0.261
d'_{asy}	3	0.000	0.107	0.509
d'_{asy}	4	0.025	0.632	0.910
d'_{asy}	5	0.018	1.966	1.652
d'_{asy}	6	1.208	3.760	2.471
T_0	1-6	279.56	338.12	33.92
s	1-5	0.0005	0.0068	0.0102
s	6	0.0019	0.0088	0.0129
R^2-adjusted		.996	.999	.999
ΔR^2		0	.001	0.0002
capacity		1	2.64	3.38

Note: the capacity was estimated by counting the number of above-threshold representations at t = 6,000 iterations.

The parameter values for the d'_{asy} are well-fitted by an exponential function, allowing the 6 free parameters to be reduced to 2 free parameters. In addition, s could be fitted with a function with only 1 parameter. Therefore, the best-fitting 9-parameter model could be further reduced to a 4-parameter model. This further parameter reduction allowed an examination of model fit as a function of differences in buffer capacity. To do this the data form the simulations with α = 1.8 and α = 2.0 were compared. This resulted in a "full" model having 8 free parameters with 4 parameters for each α-level. The 8-parameter model, [2F(d'_{asy}) – 2G(s) – 2H(T_0)], (F(x) has 2 parameters) and all nested models were fit to 120 datapoints by maximizing the adjusted R^2. Of special interest was the identification of parameters that reduce the fit and thus carry the difference in buffer capacity. The results are shown in Table 2 and are clear-cut. The goodness of fit is largely unaffected when G(s) or H(T_0) is fixed between the two levels of α. However, a 5% decrease in fit is observed when F(d'_{asy}) is fixed. The interpretation of this finding is that differences in buffer capacity are only picked up in the differences in *gradient* of the d'_{asy} function. The rate parameter seems insensitive to variation in buffer capacity and is therefore only useful to assess which item or one-chunk was the most-recently processed.

Table 2: Results of nested modeling fits on the data from the two different WM capacity simulations. The number of free parameters are given between brackets after each model.

Model	Degrees of freedom	adjusted R^2
Full model (8)	112	.989
F-fixed (6)	114	.942
G-fixed (7)	113	.988
H-fixed (7)	113	.989
F/G-fixed (5)	115	.942
F/H-fixed (5)	115	.943
G/H-fixed (6)	114	.987
All fixed (4)	116	.943

Discussion

This paper focused on the paradox that different estimates of working memory capacity are estimated based on very similar tasks. Using a dynamic model of short-term recognition, data were generated and fitted by exponential SAT functions. Contrary to what was previously thought, the results show that the rate of retrieval from WM is insensitive to the WM capacity and instead is most sensitive to the recency of cognitive processing. The asymptotic accuracy is found to be the only parameter that is sensitive to WM capacity. The resolution of the WM paradox lies in the choice of dependent measure, with accuracy being the preferred measure for estimating WM capacity and retrieval rate being the preferred measure for identifying the most recently processed chunk in WM.

The process model predicts that items that are not in WM will lead to misses. Therefore for items that were presented a very long time ago, only misses should happen. This is partially correct. One would, however, expect that deactivated items require an additional process of episodic retrieval to allow for contextual matching. This is likely to result in slower retrieval dynamics and quite likely to a larger intercept. The problem is that in order to assess this possibility, trials would have to be separated into those in which the probe matches with a deactivated item and trials in which the probe matches a pre-recency active item. This is not possible experimentally and thus differences in intercept for pre-recency items are always mixtures. The same holds for the retrieval speeds. With long lists, very early items could be probed and used to check if they do have the slowest retrieval speed and the largest intercept. The difficulty here is that performance is close to chance (Wickelgren, Corbett & Dosher, 1980). Wickelgren et al. used a 16-word list and measured the SAT of the list item -12 (position 4). In some of the participants, the intercept for the item -12 was larger than all other items. Although this might suggest that the intercept is the preferred parameter to assess whether items are retrieved from WM or form long-term memory, a thorough empirical investigation waits.

What does the reinterpretation of the exponential SAT-parameters mean for the use of the exponential SAT-procedure? Several authors have commented that exponential and diffusion SAT are too similar to be distinguished (McElree & Dosher, 1989; Ratcliff, 2006). Others have argued that diffusion SAT should be used as it is based on an actual theory of memory retrieval (Ratcliff, 2006), whereas the exponential SAT is not based on a theory and therefore only of statistically-descriptive use. Despite the finding that exponential SAT can not be used to address capacity estimates, it is able to identify the last processed item (McElree, 1998). This utility depends heavily on the assumption that across many trials, participants process the stimuli in identical ways. Whether the SAT-procedure is robust against violation of the identical-processing assumption remains for future analyses. What does all this mean for WM capacity? The analyses presented here suggest that WM can hold multiple items in an active stat to varying degrees, but that the very last processes item is in a highly accessible state. The work also demonstrates more generally the importance of using explicit formal analyses to verify the interpretations based on statistical tests.

References

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology, 49*, 5-28.

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158–73.

Broadbent, D. E. (1975). The magic number seven after fifteen year, In: *Studies in long-term memory* (Kennedy, A. and Wilkes, A., eds), pp. 3–18, John Wiley and Sons.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539-576.

Cowan, N. (1995). *Attention and memory: an integrated framework*. Oxford: Oxford University Press.

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

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*, 3-42.

Davelaar, E. J., Haarmann, H. J., Goshen-Gottstein, Y., & Usher, M. (2006). Semantic similarity dissociates short- from long-term memory: testing a neurocomputational model of list memory. *Memory & Cognition, 34*, 323-334.

Diller, D. E., Nobel, P.A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 414-435.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

Haarmann, H. J., & Usher, M. (2001). Maintenance of semantic information in capacity limited item short-term memory. *Psychonomic Bulletin & Review, 8*, 568-578.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101.

Hockley, W. E., & Murdock, B. B. (1987). A decision model for accuracy and response latency in recognition memory. *Psychological Review, 94*, 341-358.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269-299.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood appoach to the effects of experience in recognition memory. *Psychological Review, 105*, 724-760.

McElree, B. & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 18*, 346-373.

McElree, B. & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General, 122*, 291-315.

McElree, B. (1996). Accessing short-term memory with semantic and phonological information: A time-course analysis. *Memory & Cognition, 24*, 173-187.

McElree, B. (1998). Attended and non-attended states in working memory: Accessing categorized structures. *Journal of Memory & Language, 38*, 225-252.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 27*, 817-835.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation*, Vol. 46. San Diego: Academic Press.

McElree, B., & Dosher, B. A. (2001). The focus of attention across space and across time. *Behavioral and Brain Sciences, 24*, 129-130.

McKone, E. (1995). Short term implicit memory for words and nonwords. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 1108-1126

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

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review, 75*, 522-536.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review, 110*, 611-646.

Raaijmakers, J. G. W. (1982). A note on the measurement of primary memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 343-352.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of assocative memory. *Psychological Review, 88*, 93-134.

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

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*, 195-237.

Sederberg P. B., Howard M. W., & Kahana M. J. (2008) A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115, 893-912.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. (1980). Priming and retrieval from short-term memory: a speed accuracy trade-off analysis. *Journal of Verbal Learning and Verbal Behavior, 19*, 387-404.