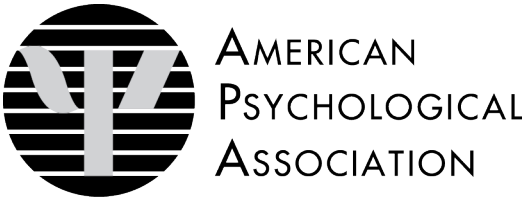

Psychological Review

Manuscript version of

When Working Memory May Be Just Working, Not Memory

Andre Beukers, Maia Hamin, Kenneth A. Norman, Jonathan D. Cohen

Funded by:

- John Templeton Foundation
- Office of Naval Research

© 2023, American Psychological Association. This manuscript is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final version of record is available via its DOI: https://dx.doi.org/10.1037/rev0000448

This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Manuscript

When Working Memory May Be Just Working, Not Memory

Andre Beukers, Maia Hamin, Kenneth A. Norman, and Jonathan D. Cohen

Department of Psychology and Princeton Neuroscience Institute

Princeton University

Author Note

Simulation code can be accessed at https://github.com/andrebeu/nback-paper. This work was supported by an award from the John Templeton Foundation to JDC and KAN and a Vannevar Bush Faculty Fellowship supported by ONR to JDC. The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the funders.

Correspondence concerning this article should be addressed to Andre Beukers, Department of Psychology, Princeton University, NJ, 08540. E-mail: abeukers@princeton.edu

Abstract

The N-back task is often considered to be a canonical example of a task that relies on working memory (WM), requiring both maintenance of representations of previously-presented stimuli and also processing of these representations. In particular, the set-size effect in this task (e.g., poorer performance on 3-back than 2-back judgments), as in others, is often interpreted as indicating that the task relies on retention and processing of information in a limited-capacity WM system. Here, we consider an alternative possibility: that retention in episodic memory (EM) rather than WM can account for both set-size and lure effects in the N-back task. Accordingly, performance in the N-back task may reflect engagement of the processing ("working") function of WM but not necessarily limits in either that processing ability nor in retention ("memory"). To demonstrate this point, we constructed a neural network model that was augmented with an EM component, but lacked any capacity to retain information across trials in WM, and trained it to perform the N-back task. We show that this model can account for the set-size and lure effects obtained in an N-back study by M. J. Kane et al. (2007), and that it does so as a result of the well-understood effects of temporal distinctiveness on EM retrieval, and the processing of this information in WM. These findings help illuminate the ways in which WM may interact with EM in the service of cognitive function, and add to a growing body of evidence that tasks commonly assumed to rely on WM may alternatively (or additionally) rely on EM.

Keywords: working memory; episodic memory; temporal context model; neural network models

When Working Memory May Be Just Working, Not Memory

1 Introduction

Immediate memory – that is, the ability to rapidly store and retrieve information after a short interval – is generally assumed to be served by two broadly distinguishable memory systems: working memory (WM) and episodic memory (EM). [1] WM is assumed to transiently maintain information in a capacity-limited fashion (Cowan, 2017; Oberauer et al., 2018). In contrast, EM stores information more durably, with few (if any) restrictions on capacity, in a latent form that can be retrieved later for use. While EM is usually not assumed to be subject to a storage capacity limitation (Polyn et al., 2009; Tulving & Thomson, 1973), retrieval from EM is subject to interference from previously studied items, referred to as proactive interference (PI). Crucially, prior work has shown that PI during retrieval can account for forgetting in tasks with short retention intervals (e.g., Brown et al., 2007; Farrell, 2012; Oberauer et al., 2012; Unsworth et al., 2011). Here, we explore this idea in the context of the N-back task, by implementing a model of the task in which the retention of information across trials relies on EM rather than WM, and processing in WM uses a neural network to compare the information retrieved from EM with the information currently represented in the network. We use this model to demonstrate that effects in the N-back task often assumed to reflect processing and maintenance constraints associated with WM function – such as the set-size and lure effects – can also be produced by PI associated with the use of EM for retention. The model provides a mechanistic grounding for recent cognitive neuroscience work addressing contributions of EM to tasks that have traditionally been construed as relying on active maintenance in WM (e.g., Beukers et al., 2021; Foster et al., 2019; Hoskin et al., 2019) and also provides a point of contact with neural network models addressing the role of EM in higher cognitive function, both within cognitive science (Webb et al., 2020) and machine learning (Graves et al., 2014; Ritter et al., 2018; Wayne et al., 2018).

WM, retroactive interference and set-size effects. WM is universally assumed to have a limited storage capacity (Oberauer et al., 2018; Sternberg, 1966), that is considered to be

[1] We specifically refrain from using the more familiar term "short-term memory" since EM, despite its capacity for rapid encoding, is usually referred to as a form of *long-term* memory due to the durability of its traces – an important factor that we discuss below.

relatively strict (in the single digits; Miller, 1956). This is often (but not always) attributed to the reliance on active maintenance as the mechanism of storage in WM, in which traces fail to be maintained because either they degrade with time, and/or are displaced by new ones. [2] The latter effect is often referred to as retroactive interference (RI; A. Baddeley, 1992; Barnes and Underwood, 1959; Peterson and Peterson, 1959). In either case, as traces decay and/or newly activated ones interfere and displace them, older information is lost. Perhaps the empirical phenomenon that best exemplifies this is the set-size effect. The set-size effect, observed across a wide range of short-term memory and WM tasks (such as the classic Sternberg paradigm, Sternberg, 1966, and the N-back task Kirchner, 1958), refers to the observation that performance degrades as more items are required to be remembered – i.e., the larger the size of the memory set, the more likely it is that information will be lost. One possible explanation of the set-size effect is that it arises from serial encoding into a limited-capacity WM system, whereby items in the set that were encoded earlier are subject to interference from those encoded later and/or decay due to the passage of time. However, the idea that set-size effects are a necessary sequela of WM engagement does not license the reverse inference, that the observation of set-size effects is a reliable indicator of WM engagement. As others have noted (e.g., Brown et al., 2007; Farrell, 2012; Oberauer et al., 2012; Unsworth et al., 2011), and we discuss below, there is a growing recognition that such effects can arise from the use of EM for storage and retrieval of recently presented information.

It is also worth noting that, in some tasks, set-size effects have also been attributed to demands on the *processing* capacity of WM which, in addition to its storage capacity, is also considered to be limited (A. D. Baddeley & Hitch, 1974). The N-back is a salient example of this (e.g., Rac-Lubashevsky and Kessler, 2016), as it has often been assumed to require the updating of the ordinal status of items in WM as each new stimulus is presented (i.e., what was the 1-back stimulus must now be assigned as the 2-back stimulus, and and the 2-back assigned as the 3-back, etc.) — a processing requirement that would obviously increase with set-size. However, as discussed below, the use of EM can avert these processing demands, while still leading to

[2] See, e.g., Oberauer et al. (2012) and Oberauer (2019) for accounts of WM that challenge this focus on active maintenance.

substantial set-size effects.

EM, proactive interference and temporal context effects. In contrast to the limited storage capacity of WM, EM is generally assumed to rely on a different mechanism of storage, in which traces are more durable (e.g., from hours to years) and are not subject to a restrictive capacity limit. However, such durable and (effectively) unrestricted storage carries with it its own limitations. Unlike WM, neither new traces nor time act to displace or degrade older memories in EM. However, as EM traces accumulate over time, the likelihood increases that a particular memory will be similar in some way to others. Because retrieval from EM is assumed to be content-based (Marr, 1971; Tulving & Thomson, 1973) – that is, items are retrieved by presenting a cue and identifying items that are most similar to it – the challenge of identifying and retrieving a particular item increases as progressively more memories are stored. One of the well-known consequences of this problem of discriminability in EM is proactive interference (PI; Brown et al., 2007): the potential for older traces in EM to be confused with newer ones that are similar, and thus interfere with reliable retrieval of the latter. Thus, although the duration and capacity of storage in EM may be unlimited, its practical use is constrained by PI at retrieval.

One important elaboration of theories concerning EM is the incorporation of temporal context information into stored traces, that can be used for later retrieval (Howard & Kahana, 2002; Lohnas et al., 2015; Polyn et al., 2009). This has been used to explain not only how people can retrieve information from particular times in the past, but also – coupled with PI – the kinds of confusion errors they make when doing so. For example, it has been used widely to explain serial position and contiguity effects in free recall tasks (Kahana, 1996, 2020) and serial recall tasks (Brown et al., 2007), many of which share similarities with tasks used to probe WM. The effects of PI associated with temporal contiguity are especially relevant in tasks that require discrimination of items presented in close temporal proximity to one another – precisely the conditions of most WM tasks, and the N-back task in particular.

Thus, despite their different properties, both WM and EM suffer from forms of interference that can constrain memory performance. Whether it is the number of items that can be retained and/or processed in WM, or the number of items that can be reliably retrieved from EM, both systems exhibit a functional limitation that can manifest as set-size effects – a

commonality that may confound the interpretation of such effects in behavioral data as evidence for the engagement of one memory system or the other.

Here, we use the N-back task to explore these possibilities, both because it is has come to be one of the most used widely probes of WM engagement (e.g., Callicott et al., 1999; Cohen et al., 1994; Dobbs and Rule, 1989; Gevins and Cutillo, 1993; Jaeggi et al., 2010; M. Kane and Conway, 2023; M. J. Kane et al., 2007; Kirchner, 1958; Nikolin et al., 2021; Oberauer, 2005; Oberauer et al., 2018; Owen et al., 2005; Rac-Lubashevsky and Kessler, 2016; Ross, 1966), and because it is generally assumed to tax both the storage and processing capabilities of WM, as reflected in the profile of performance observed in the task. Specifically, we explore an account of performance in this task that has not been widely considered, in which: the storage and retrieval of previously presented information relies exclusively on EM; WM is used only to represent and process the most recently presented stimulus and memory retrieved from EM; and processing in WM involves simply comparing and making a decision based on that information (i.e., without the need to repeatedly update which item occurred in which previous position). To the extent that this account can explain the profile of performance in the N-back, including both set-size and lure effects, then it suggests that these need not reflect constraints on storage and/or processing in WM, but rather the effects of proactive interference (PI) that can arise when stimuli with similar temporal encodings are retrieved in place of the correct ones – an effect that is consistent with temporal context models of EM and the large literature of empirical effects that are explained by these theories (Brown et al., 2007; Kahana, 2020).

We test the ability of this account to capture previously reported empirical effects in the N-back task, by implementing it in the form of a neural network model that is responsible for the representation and processing of information in WM; importantly, the model lacks any mechanism for the retention of previously presented stimuli in WM, but it is augmented with a simple form of EM that is used to encode, store and retrieve previously presented stimuli. The latter corresponds closely to a form of "external memory" (i.e., a dictionary of previous events), that is gaining increasing use as a model of episodic memory in cognitive science and neuroscience (e.g., Lu et al., 2022; Webb et al., 2020) as well as machine learning (e.g., Graves et al., 2014; Pritzel et al., 2017; Ritter et al., 2018; Wayne et al., 2018). We show that, even when there is *no*

reliance on WM for storage and the demands on processing are limited (i.e., simply making a decision based on a comparison of two sources of information), the model nevertheless exhibits empirically observed set-size and lure effects (described below), which emerge as a consequence of PI between traces in EM that incorporate similar temporal context representations.

The model consists of a feedforward neural network, coupled with a simplified implementation of a mechanism for context-based EM. The feedforward network implements the ability to compute on actively represented information (i.e., the "working" function of WM), but lacks any ability to retain that information after the relevant computations have been carried out and new information is presented to the network (i.e., it lacks the "memory" capabilities usually ascribed to WM). Rather, in the model, storage of information from one trial to the next – about the stimulus as well as temporal information that can be used to determine its serial position, both of which are required to perform the N-back task – relies on an EM module that encodes each stimulus and the temporal context in which it occurred. We first confirm that the encoding of temporal context information in EM traces causes memories encoded in close temporal proximity to interfere with one another, in a manner that can explain set-size effects observed in working memory tasks. This initial result reaffirms prior work that has demonstrated the effects of temporal distinctiveness on memory retrieval (Brown et al., 2007), here using a formally simple mechanism for temporal encoding that is consistent in its properties with previous implementations (Manning et al., 2015). We show that this mechanism, coupled with a neural network mechanism trained to evaluate the temporal "distance" between stimuli, can reproduce empirically observed patterns of performance in the N-back task (Braver et al., 1997; M. J. Kane et al., 2007).

More specifically, in the N-back task (M. Kane & Conway, 2023; Kirchner, 1958), participants see a sequence of items presented one at a time and must indicate, for each item, whether that item matches the item that occurred n items ago in the sequence. This task requires the ability to retain previously seen stimuli as well as information about their serial position, and to use that information to match the current stimulus with the relevant one retained in memory. We show that temporal context representations that change gradually with each stimulus presentation, and that are stored and can be retrieved from EM, can be used to estimate the

serial position of an earlier stimulus and thereby perform the task; however, this also makes the process subject to PI, leading to a set size effect.

In the sections that follow, we first provide an overview of the model, describing components that are relevant to all simulations. We then describe in detail how EM was implemented in the model, and discuss how similar temporal context representations can lead to PI (following Brown et al., 2007). Next we describe our implementation of the processing function of WM as a feedforward neural network, that is used to compare current information with previous information stored in EM. Finally, we use the full model to simulate performance in the N-back task, showing how – despite the absence of a mechanism for retention of prior stimuli in WM and limited demands on WM for processing, the model is able to perform the task and, in doing so, exhibits set-size effects as well as other features of human performance in the task that can be attributed to PI as a result of the use of EM for storage and retrieval.

2 Methods

2.1 Model Overview

The model consists of two components, an *EM component* and a *WM component*. The EM component is characterized by two operations: *encoding* and *retrieval*. Encoding involves storing the conjunction of features that correspond to a given stimulus. Following temporal context models of EM (Estes, 1955; Kahana, 2020), an EM trace includes stimulus features as well as the temporal context in which the stimulus occurred. Importantly, traces stored in EM are enduring (i.e., they last the entire extent of a simulation) and latent (i.e., do not influence WM processing unless retrieved). Retrieval of these latent EM traces is carried out by a similarity-based sampling operation (Gillund & Shiffrin, 1984; Graves et al., 2014; Norman & O'Reilly, 2003; Shiffrin & Steyvers, 1997; Wayne et al., 2018): The stimulus is presented on each trial as perceptual input and an associated temporal context (i.e., the current one), which are used together as a retrieval cue that is compared to all traces stored in EM; the higher the similarity between the retrieval cue and a trace stored in EM, the higher the probability of that trace being retrieved.[3]

[3] Similarity-based retrieval can be thought of as a computational approximation to the neurobiological mechanism of retrieval from EM (e.g., hippocampal pattern completion; Marr, 1971; McClelland et al., 1995). Similarity-based retrieval is also playing an increasingly important role in machine learning models that address human-level

The WM component is implemented as a strictly feedforward neural network (i.e., without any recurrence), which implements the constraint that, in this model, WM can only represent and process information that is immediately presented to it, from the environment and/or from EM; that is, it is restricted to the "working" component of WM. Specifically, its role is to compare the current perceptual information with memories retrieved from EM, and select a response based on whether the stimulus information matches while the temporal context information differs by n.

Below we show how – even though this model does not rely on WM for the retention of information across trials, and the processing demands on WM are limited (i.e., to a comparison operation but not any updating operations), while it has no constraints on the storage capacity of EM – errors can nevertheless arise due to PI that is the result of similarity in the temporal codes among traces in EM. In the following sections we describe the implementation of each of these two model components in greater detail. We start, in Section 2.2, by describing the implementation of the EM component of the model. This includes a mechanism for generating temporal context information, together with an analysis that directly examines its effects on serial position information encoded by context representations. Then, in Section 2.3 we describe the feedforward neural network architecture that is used to implement WM, and makes use of information stored in EM to perform the N-back task.

2.2 EM Component

2.2.1 EM Encoding

For every stimulus presented to the model at test, a corresponding representation was formed and stored as an EM trace. Each EM trace was a concatenation of a one-hot stimulus vector and a continuous-valued context vector (described in Section 2.2.2). Because we assume EM has no practical capacity limitation, a new trace was appended to EM storage for each stimulus that was presented over the course of a simulation. Thus, for every trial, EM contained a list of all items previously presented to the model during that simulation. Memories were encoded in EM immediately *after* being processed by the neural network as the sensory input for the

---

cognitive function, both in the attention mechanisms of transformers (Altabaa et al., 2023; Vaswani et al., 2017), and as an augmentation to neural network models with a form of external memory (e.g., Graves et al., 2014; Pritzel et al., 2017; Ritter et al., 2018; Wayne et al., 2018; Webb et al., 2020).

current trial. This was to prevent retrieval of the current stimulus from EM on the same trial in which it was also the sensory input.

2.2.2 Formulation of Context Representations

Implementation of EM in the model followed the approach taken in previous temporal context models (Estes, 1955; Howard & Kahana, 2002; Lohnas et al., 2015; Mensink & Raaijmakers, 1988; Polyn et al., 2009), using context representations that were implemented in a neurally plausible form that not only change noisily and gradually over time, but were also bounded in magnitude. To meet these criteria, we modeled context representations as an n-dimensional vector of scalar values between 0 and 1 that evolves gradually according to a random walk on an n-dimensional hypersphere. This context drift process was defined by the following equation:

$$C_t = F\left(\Phi_t^1, \Phi_t^2, \ldots \Phi_t^{n-1}\right) = F\left(\Phi_{t-1}^1 + N(\mu,\sigma), \Phi_{t-1}^2 + N(\mu,\sigma), \ldots \Phi_{t-1}^{n-1} + N(\mu,\sigma)\right) \quad (1)$$

At each timepoint, $n-1$ polar coordinates Φ_{t-1} are updated by summing a Gaussian term $N(\mu,\sigma)$.[4] That is, the context equation is defined as a Gaussian drift process on hypersphere of dimension n. We set the dimension of the hypersphere (n=25) to be sufficiently large so as to minimize the likelihood that the vector would repeat (i.e., cycle among the same set of values). We set $\mu = 0.25$ and $\sigma = 0.075$ to best fit the behavioral data. Next, in Section 2.2.3, we provide an analysis of the evolution of this temporal context representation, showing how this can give rise to PI in EM, and consequently lead to set-size effects in tasks that involve sequential presentation of stimuli over trials.

2.2.3 Analysis of Context Representation

In this section, we present analyses showing that the contextual drift (Estes, 1955; Howard & Kahana, 2002; Lohnas et al., 2015; Mensink & Raaijmakers, 1988; Polyn et al., 2009) and temporal distinctiveness (Bjork & Whitten, 1974; Brown et al., 2007; Glenberg et al., 1980) properties of our model can produce set-size effects. Then, in Section 3 we integrate this

[4] Note that to specify a hypersphere in n, only $n-1$ coordinates are needed.

mechanism with the feedforward neural network described below, and show that together these mechanisms can explain detailed patterns of behavioral performance, including set-size effects, observed for human performance in the N-back task.

Our account starts with the idea, taken from context-based models of memory, that items are tagged with a contextual representation that drifts noisily over time (Brown et al., 2007; Estes, 1955; Howard & Kahana, 2002; Lohnas et al., 2015; Mensink & Raaijmakers, 1988; Polyn et al., 2009), as described above. Contextual drift provides a basis for making temporal discriminations based on context representations. For example, consider the task of discerning the relative serial positions of two previously-presented items (e.g., which item was presented two items vs. three items ago). If the goal is to select the more recent of the two items, one approach would be to compare the context tags associated with those items to the current context and choose the item with the smaller contextual distance to the current context – the principle of contextual drift implies that on average the contextual distance should be smaller for more recent items (Hintzman, 2002). However, because of accumulated noise in the contextual drift process, the variance associated with this contextual distance measure also increases as a function of the temporal distance. As a consequence of this increase in variance, the distributions of contextual distance scores associated with nearby serial positions will overlap more as a function of elapsed time relative to some reference (e.g., the present). Concretely, if we fix the time elapsed between studying the two items the relative serial positions of which are being judged, but vary the temporal distance between these two items and the memory test (e.g., if the two items were studied in adjacent list positions, and we vary whether the two items were studied one minute ago vs. twenty minutes ago), it will be more difficult to distinguish the relative serial positions of the items as they recede into the past due to increased variance in contextual distance. This effect reflects a form of parallax, that has been described in the literature by analogy to telephone poles receding into the distance: The further the telephone poles are in the distance, the harder it is to tell apart adjacent poles (Crowder, 1976).

Previously, Brown et al. (2007) argued that the diminution of temporal distinctiveness among items as function of their distance from a reference can lead to a corresponding degradation in the ability to recall more distant items because of increased competition (i.e., if the items are

less distinguishable, it is more difficult to select out a specific item). Here, we hypothesized that this reduction in temporal distinctiveness for less recent items can provide a basis for the set-size effect in the N-back task. This effect manifests as an increase in errors and/or response times with greater *n*'s (Oberauer et al., 2018); for example, performance is worse on the 3-back version of the task than the 2-back version. As noted earlier, the set-size effect in the N-back task is often assumed to reflect a limitation in the maintenance and/or processing capabilities of WM, with 3 items subject to greater degradation or processing demands than 2 items. Here, we propose that this effect can also be produced by the increase in confusability between different context representations with increasing temporal distance, without any contribution of constraints on WM. For example, the 3-back task involves discriminating 3-back targets from items 2-back that are potential lures; and, similarly, the 2-back task involves discriminating 2-back targets from 1-back lures. By the logic outlined above, the former (3-back) task will be more difficult because the target and lures occurred further back in time, so the associated contextual distances will be more variable and thus harder to discriminate (by analogy, it is harder to determine exactly which "telephone pole" is the one 3-back); this point is illustrated in Figure 1.

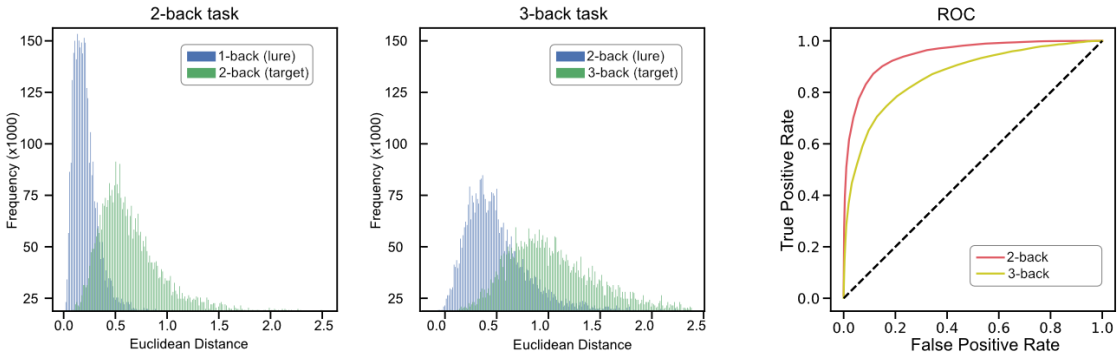

Figure 1

Lower discriminability between target item and lure for 3-back compared to 2-back. Left and middle panels: Histograms showing distance between current item and target item (green) or neighboring lure (blue) for 2-back (left) and 3-back (middle) targets. Note greater overlap between targets and lures for 3-back compared to 2-back task. Right panel: ROC for distinguishing target from lures in 3-back (yellow) and 2-back (red).

2.3 WM Component

2.3.1 Feedforward Neural Network Architecture

Processing in WM is implemented as a feedforward neural network. On each trial, the network is given information about the current task (2-back vs 3-back), the current stimulus presented to perception (letter on a computer screen), a representation of the temporal context of the current stimulus (that is distinct for each stimulus; see Section 2.2.2), and a memory trace retrieved from EM based on the current stimulus (explained below). The input layer consists of five input pools: The first two pools represent the currently-perceived stimulus (s_t; represented as a one-hot vector) and its associated temporal context representation (c_t; described below); the next two pools use the same coding scheme to represent the stimulus and context components of a memory trace (s_m, c_m, respectively), retrieved from EM (see Section 2.2); the final pool is a two-dimensional one-hot vector k that instructs the model about the task condition (e.g. 2 vs. 3 back in the N-back task). The current stimulus and context, together with the retrieved stimulus and context, are projected to the first hidden layer (h1). The one-hot task instruction vector is also projected and summed into h1. Then, the resulting vector is projected to an additional hidden layer (h2) that, in turn, projects to an output layer used to represent the response of the network to the current stimulus, indicating whether or not it judges that stimulus to match the n^{th} previously presented stimulus. Processing in the feedforward network is defined by the following equations (see also Figure 2):

$$h_1 = f\left(s_t, c_t, s_m, c_m\right) + f(k) \quad h_2 = f(h_1) \quad \text{output } = sm\left(h_2\right) \tag{2}$$

where s_t and s_m are 20-dimensional vectors representing stimuli, and c_m and c_t are 25-dimensional vectors representing context; $f(\cdot)$ is an 80 unit feedforward layer with rectified linear units (ReLU); $sm(\cdot)$ is a softmax nonlinearity mapping from hidden units to output units; and output is a 2 unit layer with activation of the first unit representing the probability of a "yes" response, and activation of the second unit representing the probability of a "no" response.

The patterns of activity provided as input to the network, together with those over its hidden units (including the representations retrieved from EM), can be thought of as the information currently represented in WM, while computations carried out on these patterns of

activity constitute the "working" function of WM. Note that none of the units in this network have persistence or integrator properties – whenever a new stimulus is presented and/or a retrieval is made from EM, they fully replace the previous corresponding patterns of activity. Nor are there any recurrent connections among units within or between layers. Thus, the network does not have the capacity to actively maintain any information in WM across time steps, nor do its prior states in any way influence its current computations. Rather, retention of information and any other effects of memories from prior trials are subserved exclusively by the EM component of the model, as described above.

2.3.2 Feedforward Neural Network Training

The network described above was trained in the simplest possible way in order to perform the N-back task. In that task, a participant is presented with a sequence of stimuli, one at a time, and must judge whether each stimulus matches the one presented n stimuli ago in the sequence. For example, in the 2-back version, for the sequence A-A-B-C-B-A the correct response is "no" for the first four stimuli and the last, while it is "yes" for the fifth stimulus (the repeat of the two-back "B"). While a naive participant might never have actually performed this particular task, people nevertheless come to the task knowing both how to match two representations based on a specified stimulus feature and how to discern relative serial positions. To capture this prior knowledge that is required to perform the task, we trained the neural network to determine whether: 1) the stimulus component of the trace retrieved from EM (s_m) was the same as the stimulus component of the current external input (s_t); and 2) the context component of the trace (c_m) was n (2 or 3) steps earlier than the context component of the current external input (c_t). That is, each training epoch consisted of a judgement about whether the stimulus and context components of a trace retrieved from EM was an n-back match to the stimulus and context components of the current input. This gave four combinations of current input and memory trace input to the network: *match* (matching stimulus, n-back context), *non-match* (non-matching stimulus, not-n-back context), *stimulus-only match* (matching stimulus, not-n-back context), and *context-only match* (non-matching stimulus, n-back context). Note that, both for simplicity and clarity of interpretation, while training the network we did not explicitly model the EM retrieval process, nor the mechanisms responsible for coordinating EM retrieval with processing by the

feedforward network.[5]. However, these generally involve the inclusion of recurrent networks that would potentially confound the interpretation of results of interest in our study. Thus, both for simplicity of implementation and clarity of interpretation, we chose not to include such mechanisms while training the present model (though we did include an EM retrieval mechanism that was coordinated with processing by the feedforward network when the model was tested as described in Section 2.4). In the General Discussion we return to this issue, which we consider an important direction for future research. The network was first pre-trained to process these four input combinations, and was then combined with the EM module to perform the N-back task (as described below).

Since training involved learning to make relative serial position determinations, and this in turn relied on the nature of the temporal context encodings (see Section 2.2.2), we trained the model on stimuli presented at various serial positions in a sequence of length 48 (the number of stimuli per block in the M. J. Kane et al., 2007 empirical study to which we compared model performance). We did so by simulating a sequence of 48 steps of temporal evolution ("drift") in the context representation (see Section 2.2.2), and then setting the context component (c_t) of the current input to a randomly selected step in that sequence. The stimulus component of the input was then selected from the set of possible one-hot stimulus vectors. The outcome of this process was to generate a sequence of stimuli that were accompanied by a drifting context representation. Finally, the task input was determined by alternating the task specification (2-back or 3-back) and randomly selecting from one of the four possible conditions, which was then used to assign the stimulus (s_m) and context (c_m) components of the trace retrieved from EM, as follows:

1. *match trial*: the same vector used for s_t was assigned to s_m ($s_t = s_m$), and c_m was the context from n steps before the current context ($c_m = c_t(t - n)$);

2. *non-match trial*: s_m was a randomly chosen one-hot vector different from the current stimulus ($s_t \neq s_m$), and c_m was randomly drawn from the context values from less than $2n$

[5] Both the mechanisms responsible for EM retrieval, and for coordinating interactions between EM and WM, are interesting and important subjects of ongoing investigation (Graves et al., 2014; Norman & O'Reilly, 2003; Pritzel et al., 2017; Ritter et al., 2018; Wayne et al., 2018; Webb et al., 2020).

steps ago excluding n ($c_m = c(t-k)$; k in $[t-(2n-1), \ldots, t-n-1, t-n+1, \ldots, t-1]$);

3. *stimulus-only match trial*: s_m was assigned as in a match trial ($s_t = s_m$), while c_m was assigned as in a "non-match" trial ($c_m = c(t-k)$; k in $[t-(2n-1), \ldots, t-n-1, t-n+1, \ldots, t-1]$);

4. *context-only match trial*: the s_m was assigned as in a non-match trial ($s_t \neq s_m$) while c_m was assigned as in a match trial ($c_m = c(t-n)$).

On each training trial, a single forward processing and backward weight-adjusting pass of the backpropagation algorithm (Rumelhart et al., 1986) was executed. The training labels were "Yes" for match trials and "No" for all other trial types. 40% of the training trials were match trials and the other 60% of training trials were evenly divided among the other trial types. Training trials were alternated between the 2-back and the 3-back conditions of the task. The model was trained on 400,000 trials per task (total of 800,000 epochs). During model testing, learning was disabled, so that no additional weight changes were possible.

2.4 Similarity-Based Retrieval, Match, and Response Processes

To simulate performance of the N-back task, we incorporated the EM mechanism described in Section 2.2, that stored traces of the stimulus and associated context (one for each item presented in a sequence of trials of the task) with the feedforward model trained on the discrimination and match process as described just above. This further required specification of how, on each trial, items were retrieved from EM and provided as input to the WM network (see note 5). Inspired by previous work that has combined neural networks with EM storage, we implemented EM retrieval using a similarity-based search process (Graves et al., 2014; Lu et al., 2022; Ritter et al., 2018; Wayne et al., 2018; Webb et al., 2020). In the present model, the current stimulus and accompanying context representation were used as retrieval cues. On each trial, the model computed the similarity of the currently presented stimulus s_t and corresponding context c_t, with each stimulus (s_m) and context (c_m) pair stored in EM. An overall memory similarity was calculated as a weighted sum of these terms:

$$sim = w_1 cos(c_t, c_m) + w_2 cos(s_t, s_m) \tag{3}$$

Note that the similarity between the current input and each memory trace was computed separately for their stimulus and context components, and weighted according to w_1, w_2 before combining them into a single similarity score for each trace. A single relative weighting of stimulus-based similarity (w_1) and context-based similarity (w_2), which was held constant across all comparisons, was optimized to best fit behavioral data (see Section 3 below). The values we arrived at were $w_1 = 0.05$ and $w_2 = 0.95$, indicating that the model was weighting context information more heavily than stimulus information when computing the similarity score. [6] The retrieval process then proceeded as follows:

1. a softmax was computed over the similarities between the current input and all traces in EM, to get the probability of retrieval of each memory;

2. the softmax values were used to probabilistically select a memory for retrieval (without replacement, as in Polyn et al., 2009; see step 5 below) which was passed to the $h1$ hidden layer of the WM network, along with the current stimulus and context;

3. if the WM network detected a match, the retrieval process terminated and the model responded "match";

4. if no match was detected in step 3, with probability $hrate$, the retrieval process was terminated and the model responded "no match" ($hrate = 0.04$ across all tasks and conditions, and was determined along with w_1 and w_2 by a fit to empirical data; see Section 3);

5. if no match was detected in step 3, and step 4 did not probabilistically trigger a "no match", steps 2-4 were repeated until the memory search terminated or until there were no more memories in EM to sample in which case the model also responded with "no match".

Thus, in summary, the EM retrieval process amounted to sampling memory traces from EM in proportion to the similarity of the currently presented stimulus and context to the stimulus

[6] This difference in weighting could potentially reflect the fact that differences in context vectors from trial to trial were smaller in magnitude than differences in stimulus vectors.

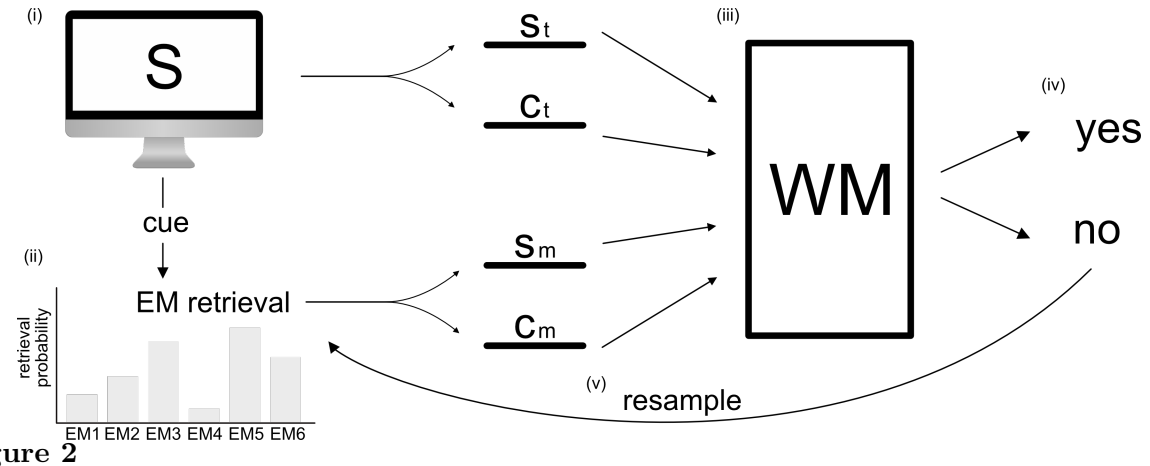

Figure 2

Model. (i) The stimulus is presented; (ii) the stimulus serves to cue memory traces; (iii) the stimulus and memory along with their respective context values are passed through the WM neural network to (iv) produce a response. If no match is found, (v) the model re-samples from EM and continues.

and context of each EM trace (weighted by w_1 and w_2, respectively), and continued on a given trial either until the retrieved trace was judged to be an N-back match to the current stimulus or it was terminated probabilistically (according to the hazard rate *hrate*).

2.5 N-Back Simulation and Analysis

We simulated the experiments conducted by M. J. Kane et al. (2007), which compared 3-back versus 2-back in 8 blocks of 48 trials each, using 8 phonologically distinct letters. The analysis involved distinguishing eight different conditions, defined by the crossing of three factors: set size (2- vs. 3-back instruction), match vs. non-match (does the current stimulus match the n-back stimulus), and the presence or absence of a lure (does the current stimulus also match the n-1 back stimulus). Crossing the match and lure factors yielded four sequence types; for example, in the 3-back condition, these were: (i) match sequences (A B C A); (ii) non-match sequences (B C D A); (iii) match-lure sequences (A A B A); and (iv) non-match-lure sequences (B A C A). [7] Thus, sampling these four sequence types for the two set sizes (2-back and 3-back) yielded eight

[7] In M. J. Kane et al. (2007), these sequence types are referred to as *control target*, *control foil*, *lure target*, and *lure foil*, respectively.

conditions, that we used in our simulations. Following the M. J. Kane et al. (2007) study, simulated trials were drawn from blocks of 48; each trial involved presenting a stimulus drawn from one of the eight possible conditions at a particular point in the block, and the model had to judge whether that stimulus was an n-back match or not. To simulate the t^{th} trial in a block, the current (t^{th}) stimulus was randomly drawn from the set of possible stimuli (e.g., it was set to A). Then, depending on the experimental condition being simulated, the preceding n stimuli were selected to instantiate that condition, by loading them into EM. To simulate the key assumption that EM traces are durable, EM was also loaded with randomly selected stimuli for all of the trials in that block preceding the one n-back. For example, in the 19^{th} trial of a 3-back block, if the current stimulus was A in a match sequence, then the 16^{th}, 17^{th}, and 18^{th} stimuli would be chosen as A, B, and C, so that 3-back stimulus (A) matched but the 2-back (B) did not. Then we filled in the stimuli for the preceding trials (i.e., the 1^{st} through 15^{th} trials) randomly from the set of all possible stimuli, such that on the 19^{th} trial, the model has 18 EM traces available for retrieval, some of which could match the current stimulus. Finally, we generated a drifting sequence of t context vectors (one per trial), in which the t^{th} context vector in the sequence was designated as the current context, and we stored episodic memories for the $t-1$ trials preceding the current trial each paired with a corresponding context vector (e.g., the EM trace corresponding to the 15^{th} trial would contain the 15^{th} stimulus and the 15^{th} context vector). Once the current stimulus and context were fixed, and the contents of EM were defined, the model produced a response as described in Section 2.4.

3 Results

We compared the results of the simulations described above with those reported by M. J. Kane et al. (2007), using the same signal-processing metrics to analyze performance that they used to analyze their empirical data (hits, correct rejections, d' sensitivity, and C bias). These were calculated separately for each of the eight conditions (again, 2-back and 3-back set sizes crossed with the four sequence types). Figure 3 shows that the pattern of results from the simulations closely matched those of the empirical study.

Set-size effect. In the N-back task, the set-size effect manifests as lower sensitivity in the 3-back task compared to the 2-back task. This effect was robustly present both in humans and

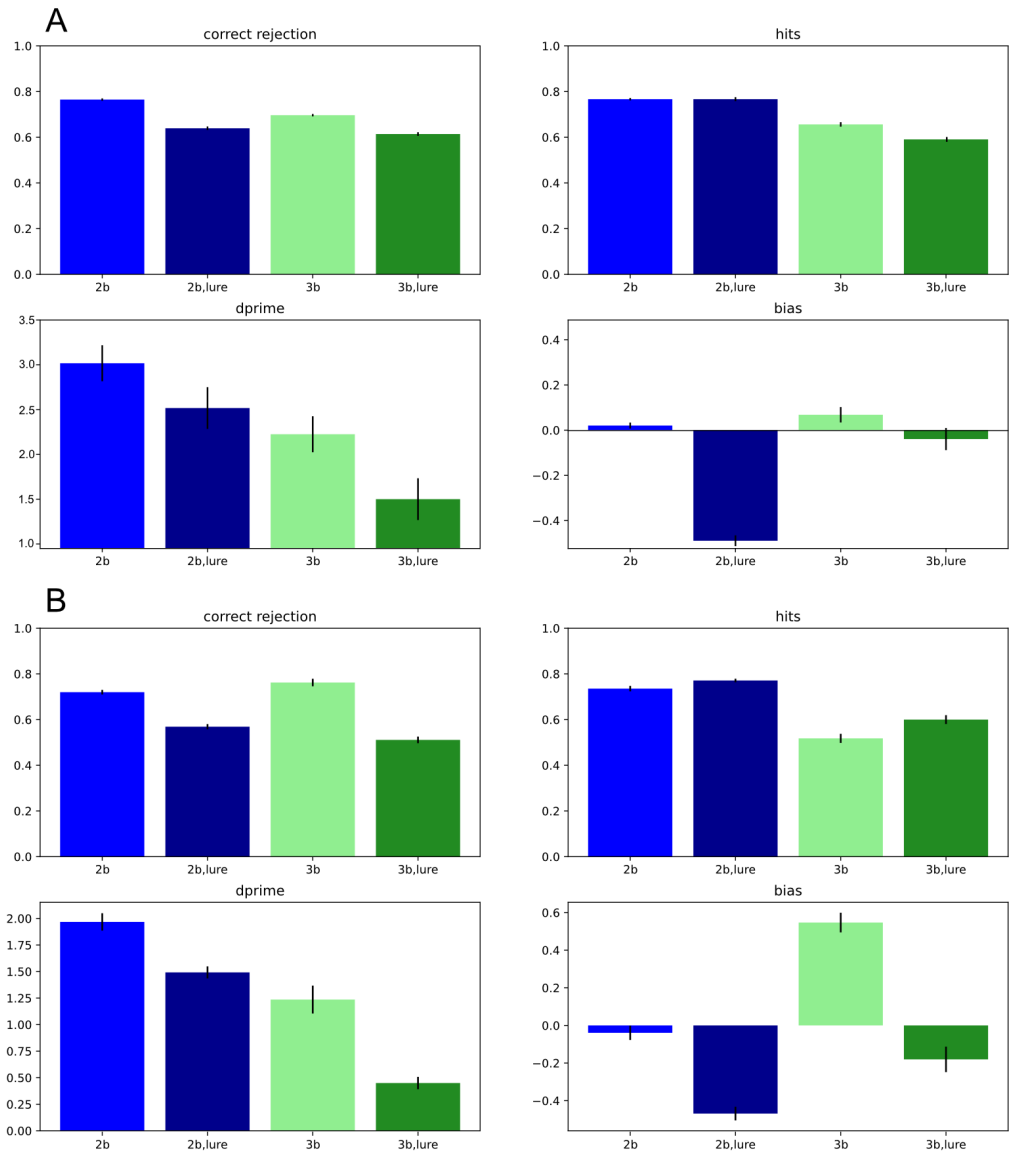

Figure 3

A) Human and B) model results, showing hit rate (correctly replying "match" on match and match-lure trials), correct rejection rate (correctly replying "no match" on non-match and non-match-lure trials), sensitivity, and bias, as a function of set size (2-back vs. 3-back) and whether or not a lure was present at the n-1-back position (see text for explanation of sequence types). Error bars indicate the standard error of the mean. A) Human data reproduced from M. J. Kane et al. (2007). B) Data from model simulations, averaging across 10 runs of the model, corresponding to N=10 participants.

the model. The human data shows a strong main effect of lower hit-rate in the 3-back compared to the 2-back condition. This was also observed in the model. [8]

Lure effects. In the N-back task modeled here, we use the term "lure" to refer to a sequence of stimuli in which a stimulus in the n-1 position matches the current stimulus. This could occur either in match sequences (which also contain an n-back match) or in non-match sequences (which do not contain an n-back match). The lure effect manifests as lower sensitivity in conditions with (vs. without) lures, as the n-1 back item is likely to be confused with the n-back item. We found that this effect of reduced sensitivity was robustly present both in humans and in the model. The lure effect was primarily driven by higher false alarm rates to non-match-lure sequences (e.g., for 3-back, B A C A) than to non-match sequences (B C D A). False alarms to non-match-lure sequences occur because noise accumulation in context drift can lead to confusion between adjacent serial positions (see Figure 1); for example, in the sequence B A C A, when the A that occurred in the n-1 position is retrieved, the retrieved context is sometimes mis-attributed to the n-back position, leading to a spurious "match" response. Interestingly, the effect of the lure manipulation on hits was different in the model as compared to humans. The model showed more hits on lure trials than non-match trials for the same reason it showed more false alarms on lure trials – it would sometimes give a spurious "match" response after retrieving a lure). Humans showed the opposite pattern of results in the 3-back condition (more hits on control trials than lure trials), an observation that will require additional modeling and/or empirical work to understand.

4 General Discussion

In this article, we presented a model of the N-back task that simulated human performance on this task, and exhibited set-size effects that arose strictly from PI due to contextual drift and degradation of temporal distinctiveness in the retrieval of information from EM. Our findings lend support to the view that EM may be engaged by – and contribute to set-size effects in – tasks widely used to index WM function, as discussed further below. The demonstration of these effects is particularly relevant in the context of the N-back task, given

[8] We report findings here for the two set-sizes studied in M. J. Kane et al. (2007). Findings for a wider range of set-sizes are reported in the Appendix.

both its wide use as an index of WM function (M. J. Kane et al., 2007; Oberauer et al., 2018), and because our findings suggest that the strong set-size effects observed in this task need not be attributed to constraints in either the storage *or* processing capacity of WM (e.g., Rac-Lubashevsky and Kessler, 2016). As reviewed above, our model showed robust set-size effects despite having no capacity whatsoever for the retention of information in WM (all information about previous stimuli was stored in EM), and despite having sufficient WM processing capacity to handle the demands of the task (WM processing simply involved comparing the information retrieved from EM with the information currently in WM and generating a response). In the remainder of this discussion, we consider the relationship of our model to other models of the N-back task, as well as current theories of EM, WM, and their interactions.

4.1 Relationship to Existing N-back Models

To date, there have been relatively few published mechanistic models of the N-back task. Here we compare our model to two of these that are representative of how previous work has treated the role of WM in performance on the N-back task. One of these models, reported by Chatham et al. (2011), used a neural network to implement a biologically-plausible mechanism for the active maintenance and processing of information in WM, based on a previous model of prefrontal cortex and basal ganglia function (Frank et al., 2001). Chatham et al. (2011) showed that this model could replicate set size and lure effects in the N-back task. Our findings complement these results, showing that a neural network model that uses WM to process information (i.e., evaluate for a match and elicit a response), but that relies exclusively on EM rather than active maintenance in WM for retaining information across trials, can produce comparable results.

Another model, reported by Juvina and Taatgen (2007), explores how two different strategies can be used to perform the N-back task. One of these relies on the active maintenance of information in WM, paralleling at an abstract level the Chatham et al. (2011) model. The other strategy relies on a form of storage similar in important respects to more recently proposed alternative forms of storage in WM (Oberauer, 2019; Stokes, 2015) (discussed further below, in Section 4.3). Juvina and Taatgen refer to these strategies as “high control” and “low control”, respectively, and implemented them in two distinct models using the ACT-R architecture

(Anderson et al., 1997). In their high-control model, a window of size of n stimuli was actively maintained by a rehearsal process, and the ordinal position of each item was encoded by the item's position in this actively-maintained window. This can be thought of as implementing a WM-based mechanism for retaining information in an actively-maintained state. In contrast, in their low-control model, each item was stored along with a time-tag that specified the moment of encoding, following the time-tag account of Yntema and Trask (1963). This implements a form of temporal context dependence similar to the one implemented in our model: time-tags were encoded in memory and then retrieved (in response to repeated stimuli) and used to make serial position judgments.

Critically, however, an important difference between the Juvina and Taatgen's low-control model and the one presented here (aside from our use of a neural network to perform the match and response processes), is that their model relied on a memory decay mechanism to explain the set-size effect (i.e., in their model, memories decay with time, making 3-back targets less likely to be retrieved than 2-back targets). By contrast, the model presented here relied on noisy contextual drift and temporal distinctiveness to explain set-size effects, without positing any dedicated decay mechanisms. This reliance on temporal distinctiveness (and not decay) to explain set-size effects aligns with classic work suggesting that decay, on its own, is not a major source of forgetting in episodic memory (e.g., A. Baddeley and Hitch, 1977). As discussed in the next section, our model implements time-tags in a form that is also closely related to other models of temporal context based memory (Howard & Kahana, 2002; Polyn et al., 2009), while using a neurally plausible representational coding scheme (i.e, as value-constrained drifting context vectors), and shows how a simple neural network model can learn to use such context vectors to perform the temporal discrimination and matching processes required by the task. Thus, while our model aligns with the theoretical proposition advanced in Juvina and Taatgen (2007) and by others (e.g., Oberauer et al., 2012) – that set-size effects may not necessarily reflect reliance on active maintenance in WM for retention – it relies on a different mechanism for explaining degradation in performance with set size, which is more closely aligned with contemporary work on storage and retrieval from EM (discussed below) than other forms of WM, while also offering a neurally-plausible implementation of the mechanisms involved. It also relates closely to the

growing body of work on neural network models that make use of interactions between EM and WM for higher cognitive functions, such as planning and generalization, to which we return in Section 4.4.

4.2 Relationship to Existing Context and Temporal Distinctiveness Models

Our model of storage and retrieval from EM aligns closely with existing context-based memory models, such as the Temporal Context Model (TCM; Howard and Kahana, 2002) and the Context Maintenance and Retrieval model (CMR; Lohnas et al., 2015; Polyn et al., 2009). These models explain a wide range of findings from serial recall and free recall paradigms in terms of a gradually-drifting temporal context representation. For example, recency effects in free recall (i.e., better recall of more recent items) can be explained as a consequence of a greater match between the current context and the context associated with recent (vs. more temporally distant) memories (Howard & Kahana, 2002). Here, we focus on a different consequence of the similarity properties of context representations as a function of time: If drifting context representations carry serial position information that is reinstated by retrieved EM traces, this serial position information can be used for carrying out task-relevant computations in WM – in this case, identifying whether an item was presented 2-back or 3-back. Extending the work of Brown et al. (2007) and others, we show how confusability of retrieved context representations can lead to memory errors in the N-back task.

4.3 Relationship to "Dual System" Models of Immediate Memory

The model we present here is certainly not the first to posit that more than one system may contribute to immediate memory. Dating back to at least James (1890), it has been acknowledged that such memory is best explained as multiple interacting systems. Here we compare and contrast our model with two notable instances of such models in the literature: the dual system framework of Unsworth and Engle (2007) and the SOB-CS model of Oberauer et al. (2012).

Following from terminology introduced by William James, Unsworth and Engle (2007) describe a model composed of two interacting memory systems, primary memory and secondary memory. Similar to traditional concepts of WM, primary memory is described as a dynamic and attention-driven component that manipulates a small number of items (2-7). This capacity

constraint is then used to explain limitations in performance such as the set-size effect. Unsworth and Engle (2007) also describe a secondary memory component that, like our EM module, has no capacity constraints and from which items must be retrieved by probabilistic cue-dependent retrieval process. The difference between this secondary memory mechanism and the model of EM presented here relates to the process responsible for retrieval. Specifically, retrieval from secondary memory is achieved through an active search process, that strategically formulates cues to delineate a search set from which a memory is sampled. In contrast, in our model retrieval from EM is automatically triggered on every trial, by the similarity between the item and context currently active in WM state and the item and context stored in candidate EM traces. Thus, while the framework in which our model was constructed allows for the possibility that additional strategic processes may be involved in actively searching for and selecting cues to constrain what information is retrieved from EM (e.g., processes that might come into play at the very beginning of the experiment, to determine that the the current stimulus and context are all that are needed to cue retrieval from EM), our model suggests that such strategic mechanisms are not needed – at least not on a trial-by-trial basis – to account for the set-size and lure effects that are observed empirically.

Along these lines, an additional parallel between our work and that of Unsworth and Engle (2007) relates to the use of a context representation to guide the retrieval process. Although the coding scheme of the context representation is not formally specified by their theory, the authors make the intriguing suggestion that context could be hierarchically specified. For example, when studying a list of items, each item would be associated with a global context, a list context, and an item-level context. Here we formally represent the item-level context as an automatic drift process, in line with existing formalisms from the EM literature, and in particular the temporal context memory model Howard and Kahana (2002). An interesting future direction might be to explore what other phenomena could be modeled if we allowed context representations to be hierarchically structured, and/or have different dynamics under different circumstances that might even be strategically controlled. For example, to model the effect of list-level context, we could either augment the context vector to have different entries that drift at different rates, and/or allow for the context to take a single large step at the end of a list. In both cases, this should

have the effect of making items from different lists more distinguishable than ones within lists.

In another dual-system model, Oberauer et al. (2012) built on the "context-serial-order-in-a-box" (C-SOB) model (Farrell, 2006; Lewandowsky & Farrell, 2008) to propose the SOB-CS model for the complex span task. On each trial of this task, participants are given a list of items to remember followed by a distractor task (e.g., doing algebra, or making a lexical decision) after which they must recall the list. The task is designed to tax the participant's ability to retain information while performing manipulations in working memory. Like the Unsworth and Engle (2007) model, the SOB-CS model is also composed of two components. Similar to primary memory, SOB-CS has a mechanism for the focus of attention that contains a limited amount of activated information available for processing (corresponding to WM). Similar to secondary memory, memories that are in the focus of attention are stored along with their context in a more durable component from which items can later be retrieved (corresponding to EM).

Our model is similar to SOB-CS in that it explains memory failures in terms of interference (as opposed to decay) mechanisms. However, it differs with respect to how this arises. In SOB-CS, items are associated with their respective context by Hebbian learning in a weight matrix, which is used to store them in secondary memory. Importantly, this applies both to memory list items and items in the distractor task. Because memories are stored in the same weight matrix that has a fixed dimensionality (i.e., constrained capacity), they can interfere with and eventually start overwriting each other. Oberauer et al. (2012) show that, as memories accumulate in the memory matrix, such interference can explain a series of memory effects including the set-size effect and serial position effects in list recall. The effects in our model have some similarities to those exhibited by SOB-CS, and at some level of abstraction these models may be formally related to one another. However, the mechanisms underlying these effects differ: in our model memories in EM accumulate independently without directly interfering with one another, and interference is driven entirely by the effects of similarity structure on retrieval – in particular, the similarity of items with respect to their temporal context representations.

Another difference between our model and SOB-CS is the forgetting mechanism. To prevent interference from completely bogging down memory, Oberauer et al. (2012) posit an

active retrieval mechanism implemented by Hebbian anti-learning that clears the irrelevant items from the memory matrix during free time. In contrast, in our model, the "clearing effect" does not require any dedicated machinery; rather, once again it emerges directly from an interaction between the two fundamental mechanisms in the model, encoding of slowly drifting temporal context representation and similarity-based retrieval: As the current temporal context representation drifts further away from the context representations associated with previous items, those previous items become less accessible for retrieval and thus less interfering.

Perhaps the most important, broader observation to be made about efforts to model immediate memory is that the gap between models of WM and EM has narrowed considerably: Mechanisms that previously had been part only of EM models (e.g., item-context binding) are now frequently included in models of WM, such as the Oberauer et al. (2012) model discussed above; at the same time, effects historically associated with WM (such as set-size effects) are increasingly being considered with respect to EM. In this context, we emphasize that the goal of the work reported here was not to promote the exclusive, or even primary role of EM in contributing to retention of information in WM tasks (in general) and the N-back task (in particular). Rather, it was to help refine the functional definition of and distinction between these memory mechanisms, and to demonstrate in as clear a way as possible that three fundamental properties associated with EM (similarity-based retrieval of durable memory traces that bind items to a drifting temporal context) are *sufficient* to account for set-size and lure effects in a task widely used to probe working memory function (i.e., the N-back task), without implying either their necessity or primacy.

4.4 Other Interactions between EM and WM

In this theoretical note, for the reasons just mentioned, we focused on one form of complementary interaction between EM and WM, in which EM serves as mechanism for retention, and WM for computation. In reality, it is almost certain that performance in any given task relies on the participation of both, in ways that vary by task condition. Indeed, a growing number of models suggest interactions between EM and WM may be a central, not just an incidental feature of cognitive function (A. Baddeley, 2000; Beukers et al., 2021; Cowan, 1999, 2019; Dulberg et al., 2021; Foster et al., 2019; Oberauer, 2009; Rose, 2020; Webb et al., 2020). For example, Cohen and

O'Reilly (1996) proposed that an interaction between these systems may provide an account of prospective memory – that is, remembering to perform a task in the future (Einstein & McDaniel, 2005) – in which episodic memory serves to store an association between a representation of the desired task and the circumstances in which it is to be performed, so that the task representation can be retrieved when those circumstances occur. This provides a mechanistic undergirding of two-process theories of prospective memory (Einstein & McDaniel, 2005; McDaniel & Einstein, 2000, 2007), which have received empirical support from both behavioral and neural data (e.g., Beck et al., 2014; Einstein et al., 2005; Lewis-Peacock et al., 2016; McDaniel et al., 2013), and have recently been subjected to normative analysis (Momennejad et al., 2021). Furthermore, recent work in machine learning has suggested that interactions between EM and recurrent neural network mechanisms that support gradual learning and WM (such as long short-term memory mechanisms, LSTMs; Hochreiter and Schmidhuber, 1997) may be critical for other forms of higher cognitive function, such as the learning of abstract rules and their use in reasoning and problem solving (Altabaa et al., 2023; Graves et al., 2014; Vaishnav & Serre, 2022; Webb et al., 2020). In this light, the work presented in this article contributes to research addressing the relationship between EM and WM, a direction that promises to be an increasingly important and productive one for understanding higher cognitive function.

4.5 Concluding Remarks

Advancing our understanding of human cognitive function requires understanding how different subsystems interact, including the different memory systems. In this theoretical note we focused on one potential interaction, between WM and EM. Our model provides a neural network implementation of how context representations may be encoded in and used for retrieval from EM, and how this may be used for WM computations involving serial position information. We describe a computational account of these interactions, showing that these interactions can produce set-size and lure effects in the N-back task similar to those observed in human performance. The robust nature of these effects in the N-back task has likely contributed to its widespread use as an index of WM function, under the assumption that they reflect limits to the storage and/or processing capacity commonly ascribed to WM. The work presented here joins other lines of work that suggest caution is warranted in this inference, reinforcing the idea that

well-established properties of EM – similarity-based retrieval of temporal context representations – are sufficient to elicit such effects. Future work should focus on the development of more detailed experimental and analysis methods that can be used to disambiguate the contributions of WM and EM to task performance, as well as their interaction in the service of memory and their role in higher cognitive function. We also hope that the formulation of our model within the context of a neural network architecture will facilitate contact with work both in neuroscience and machine learning that explores the mechanisms of interaction between EM and WM.

5 Appendix

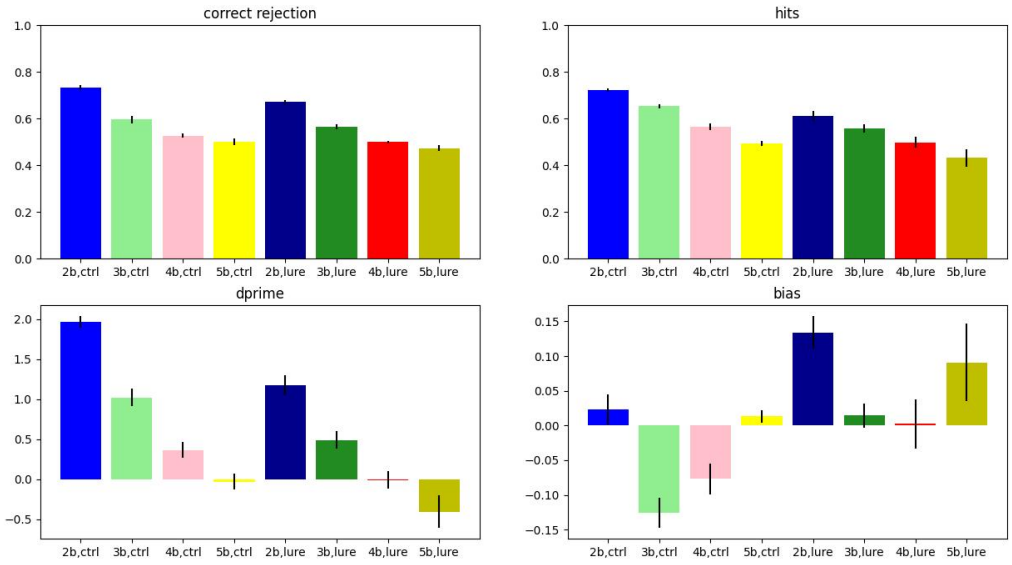

Figure 4

Extended version of model results from Figure 3 showing hit rate (correctly replying "match" on match and match-lure trials), correct rejection rate (correctly replying "no match" on non-match and non-match-lure trials), sensitivity, and bias, as a function of set size (2-back all the way through 5-back) and whether or not a lure was present at the n-1-back position.

Figure 4 shows an extended version of the model results from Figure 3, with N-back values extending from 2 to 5. Methods for these simulations were the same as for the simulations in the main text; the only difference is that 4-back and 5-back conditions were included. Sensitivity (indexed using d prime) drops to zero in the 5-back control condition and the 4-back lure

condition. This finding could naively be interpreted as reflecting a strict limit in storage and/or processing capacity, but – in the model – it is driven by the decreasing discriminability of retrieved contexts for adjacent serial positions as N-back increases.

References

Altabaa, A., Webb, T., Cohen, J., & Lafferty, J. (2023). Abstractors: Transformer modules for symbolic message passing and relational reasoning. *arXiv preprint arXiv:2304.00195*.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction*, *12*(4), 439–462. https://doi.org/10.1207/s15327051hci1204_5

Baddeley, A. (1992). Working memory [Publisher: American Association for the Advancement of Science Section: Articles]. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A., & Hitch, G. J. (1977). Recency re-examined. In *Attention and Performance VI* (S. Dornic (Ed.)). Lawrence Erlbaum Associates.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation*, *8*, 47–89.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*(2), 97–105. https://doi.org/10.1037/h0047507

Beck, S. M., Ruge, H., Walser, M., & Goschke, T. (2014). The functional neuroanatomy of spontaneous retrieval and strategic monitoring of delayed intentions. *Neuropsychologia*, *52*, 37–50.

Beukers, A. O., Buschman, T. J., Cohen, J. D., & Norman, K. A. (2021). Is activity silent working memory simply episodic memory? [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *25*(4), 284–293. https://doi.org/10.1016/j.tics.2021.01.003

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall [Place: Netherlands Publisher: Elsevier Science]. *Cognitive Psychology*, *6*(2), 173–189. https://doi.org/10.1016/0010-0285(74)90009-7

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*(1), 49–62.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. https://doi.org/10.1037/0033-295X.114.3.539

Callicott, J. H., Mattay, V. S., Bertolino, A., Finn, K., Coppola, R., Frank, J. A., Goldberg, T. E., & Weinberger, D. R. (1999). Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cerebral Cortex*, *9*(1), 20–26.

Chatham, C. H., Herd, S. A., Brant, A. M., Hazy, T. E., Miyake, A., O'Reilly, R., & Friedman, N. P. (2011). From an executive network to executive control: A computational model of the *n*-back task. *Journal of Cognitive Neuroscience*, *23*(11), 3598–3619. https://doi.org/10.1162/jocn_a_00047

Cohen, J. D., Forman, S. D., Braver, T. S., Casey, B., Servan-Schreiber, D., & Noll, D. C. (1994). Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI. *Human Brain Mapping*, *1*(4), 293–304.

Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In *Prospective memory: Theory and applications* (pp. 267–295). Lawrence Erlbaum Associates Publishers.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N. (2019). Short-term memory based on activated long-term memory: A review in response to norris (2017). *Psychological Bulletin*.

Crowder, R. G. (1976). *Principles of learning and memory*. Lawrence Erlbaum.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and aging*, *4*(4), 500.

Dulberg, Z., Webb, T., & Cohen, J. (2021). Modelling the development of counting with memory-augmented neural networks. *arXiv preprint arXiv:2105.10577*.

Einstein, G. O., & McDaniel, M. A. (2005). Prospective memory: Multiple retrieval processes. *Current Directions in Psychological Science*, *14*(6), 286–290. https://doi.org/10.1111/j.0963-7214.2005.00382.x

Einstein, G. O., McDaniel, M. A., Thomas, R., Mayfield, S., Shank, H., Morrisette, N., & Breneiser, J. (2005). Multiple processes in prospective memory retrieval: Factors determining monitoring versus spontaneous retrieval. *Journal of Experimental Psychology: General*, *134*(3), 327.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression [Place: US Publisher: American Psychological Association]. *Psychological Review*, *62*(3), 145–154. https://doi.org/10.1037/h0048509

Farrell, S. (2006). Mixed-list phonological similarity effects in delayed serial recall. *Journal of Memory and Language*, *55*(4), 587–600. https://doi.org/10.1016/j.jml.2006.06.002

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological review*, *119*(2), 223.

Foster, J. J., Vogel, E. K., & Awh, E. (2019). Working memory as persistent neural activity. *PsyArXiv*. https://doi.org/10.31234/osf.io/jh6e3

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective & Behavioral Neuroscience*, *1*(2), 137–160. https://doi.org/10.3758/cabn.1.2.137

Gevins, A., & Cutillo, B. (1993). Spatiotemporal dynamics of component processes in human working memory. *Electroencephalography and clinical Neurophysiology*, *87*(3), 128–143.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67. https://doi.org/10.1037/0033-295X.91.1.1

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., Gretz, A. L., Fish, J. H., & Turpin, B. M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(4), 355–369. https://doi.org/10.1037/0278-7393.6.4.355

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines [arXiv: 1410.5401]. *arXiv:1410.5401 [cs]*. Retrieved August 3, 2020, from http://arxiv.org/abs/1410.5401

Hintzman, D. L. (2002). Context matching and judgments of recency. *Psychonomic Bulletin & Review*, *9*(2), 368–374. https://doi.org/10.3758/BF03196295

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory [Publisher: MIT Press]. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoskin, A. N., Bornstein, A. M., Norman, K. A., & Cohen, J. D. (2019). Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience*, *19*, 338–354.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the n-back task as a working memory measure. *Memory*, *18*(4), 394–412.

Juvina, I., & Taatgen, N. A. (2007). Modeling control strategies in the n-back task. *Proceedings of the 8th International Conference on Cognitive Modeling*, 73–78.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109. https://doi.org/10.3758/BF03197276

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(1), 107–138. https://doi.org/10.1146/annurev-psych-010418-103358

Kane, M., & Conway, A. (2023). The invention of n-back: An extremely brief history. *Authorea Preprints*.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622. https://doi.org/10.1037/0278-7393.33.3.615

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Lewandowsky, S., & Farrell, S. (2008). Short-term memory: New data and a model. *Psychology of Learning and Motivation*, *49*, 1–48.

Lewis-Peacock, J. A., Cohen, J. D., & Norman, K. A. (2016). Neural evidence of the strategic choice between working memory and episodic memory in prospective remembering. *Neuropsychologia*, *93*, 280–288.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological review*, *122*(2), 337.

Lu, Q., Hasson, U., & Norman, K. A. (2022). A neural network model of when to retrieve and encode episodic memories [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *11*, e74445. https://doi.org/10.7554/eLife.74445

Manning, J., Norman, K., & Kahana, M. (2015). The role of context in episodic memory. In M. Gazzaniga & R. Mangun (Eds.), *The Cognitive Neurosciences V*. Cambridge, MA: MIT Press.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *262*(841), 23–81. https://doi.org/10.1098/rstb.1971.0078

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*.

McDaniel, M. A., & Einstein, G. O. (2000). Strategic and automatic processes in prospective memory retrieval: A multiprocess framework. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *14*(7), S127–S144.

McDaniel, M. A., & Einstein, G. O. (2007). Spontaneous retrieval in prospective memory. *The foundations of remembering: Essays in honor of Henry L. Roediger, III*, 225–240.

McDaniel, M. A., LaMontagne, P., Beck, S. M., Scullin, M. K., & Braver, T. S. (2013). Dissociable neural routes to successful prospective memory. *Psychological science*, *24*(9), 1791–1800.

Mensink, G.-J., & Raaijmakers, J. G. (1988). A model for interference and forgetting [Place: US Publisher: American Psychological Association]. *Psychological Review*, *95*(4), 434–455. https://doi.org/10.1037/0033-295X.95.4.434

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2). https://doi.apa.org/doiLanding?doi=10.1037%2Fh0043158

Momennejad, I., Lewis-Peacock, J., Norman, K. A., Cohen, J. D., Singh, S., & Lewis, R. L. (2021). Rational use of episodic and working memory: A normative account of prospective memory. *Neuropsychologia*, *158*, 107657.

Nikolin, S., Tan, Y. Y., Schwaab, A., Moffa, A., Loo, C. K., & Martin, D. (2021). An investigation of working memory deficits in depression using the n-back task: A systematic review and meta-analysis. *Journal of Affective Disorders*, *284*, 1–8.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646. https://doi.org/10.1037/0033-295X.110.4.611

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of experimental psychology: General*, *134*(3), 368.

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation*, *51*, 45–100.

Oberauer, K. (2019). Working memory capacity limits memory for bindings. *Journal of Cognition*, *2*(1), 40. https://doi.org/10.5334/joc.86

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198. https://doi.org/10.1037/h0049234

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., Wierstra, D., & Blundell, C. (2017). Neural episodic control. *International conference on machine learning*, 2827–2836.

Rac-Lubashevsky, R., & Kessler, Y. (2016). Decomposing the n-back task: An individual differences study using the reference-back paradigm. *Neuropsychologia*, *90*, 190–199.

Ritter, S., Wang, J., Kurth-Nelson, Z., & Botvinick, M. (2018). Episodic control as meta-reinforcement learning. *arXiv*. https://doi.org/10.1101/360537

Rose, N. S. (2020). The dynamic-processing model of working memory. *Current Directions in Psychological Science*, *29*(4), 378–387.

Ross, B. M. (1966). Serial order as a unique source of error in running memory. *Perceptual and Motor Skills*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Sternberg, S. (1966). High-Speed Scanning in Human Memory [Publisher: American Association for the Advancement of Science]. *Science*, *153*(3736), 652–654. Retrieved March 1, 2022, from https://www.jstor.org/stable/1719418

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071