Frequency effects on memory: A resource-limited theory

Vencislav Popov (vpopov@cmu.edu)[1,2] & Lynne Reder (reder@cmu.edu)[1,2]

[1] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[2] Center for the Neural Basis of Cognition, Pittsburgh, PA

Accepted for publication in Psychological Review. The final published version might differ from the current manuscript.

Abstract

We present a review of frequency effects in memory, accompanied by a theory of memory, according to which the storage of new information in long-term memory (LTM) depletes a limited pool of working memory (WM) resources as an inverse function of item strength. We support the theory by showing that items with stronger representations in LTM (e.g. high frequency items) are easier to store, bind to context, and bind to one another; that WM resources are involved in storage and retrieval from LTM; that WM performance is better for stronger, more familiar stimuli. We present a novel analysis of preceding item strength, in which we show from nine existing studies that memory for an item is higher if during study it was preceded by a stronger item (e.g. a high frequency word). This effect is cumulative (the more prior items are of high frequency, the better), continuous (memory proportional to word frequency of preceding item), interacts with current item strength (larger for weaker items) and interacts with lag (decreases as the lag between the current and prior study item increases). A computational model that implements the theory is presented, which accounts for these effects. We discuss related phenomena that the model/theory can explain.

Contents

I. Introduction

When we learn new information, we have to combine or chunk multiple features of our experiences into a single whole. The building blocks of these experiences can be, for example, faces, names, concepts, individual sensations, or an experiential context. We can think of these units and the associations among them as traces in a memory system, and some of these traces are arguably stronger than others – people find it much easier to recall and use some words, concepts, or events, compared to others. When it comes to memory trace strength, there are a number of questions which any theory of memory has to answer: How should we conceptualize memory strength? What are the factors that make some memories stronger than others and what are the mechanisms through which changes in memory strength occur? How and why does the existing strength of a memory affect future learning?

Much of memory research revolves around the commonsense concept of memory trace strength[1], yet there is little agreement about how to operationalize it. Some models in the field consider strength to be an intrinsic property stored within memory traces themselves (e.g., Anderson, Matessa, & Lebiere, 1997; Reder et al., 2000), while others calculate strength at retrieval, a measure of evidence for the match between a cue and the contents of memory (e.g. Dennis & Humphreys, 2001; Hintzman, 1984; Murdock, 1982; Osth & Dennis, 2015; Shiffrin & Steyvers, 1997). While both classes of models can account for many key findings in the field, they significantly disagree on whether many of the effects reviewed in this paper occur at encoding or at retrieval. Relatedly, memory theorists have long debated whether memory traces are best described by a single strength value or whether multiple signals contribute to memory judgements (Anderson & Bower, 1972; Diana, Reder, Arndt, & Park, 2006; Wixted & Mickes, 2010; Yonelinas, 2002).

In operationalizing memory strength, theorists have to deal with a number of intriguing behavioral patterns. The prior strength of items in memory has multifaceted and often opposing effects on memory encoding, binding and retrieval (for an earlier review, see Reder, Paynter, Diana, Ngiam, & Dickison, 2007). Depending on the nature of the task, the procedure, the experimental design, the stimuli and the processes under investigation, prior item strength can either facilitate or impair memory performance. Normative word frequency has been a favorite stimulus variable in service of our understanding of the effect of prior item strength on memory. A myriad of studies indicate that high frequency words do worse than low frequency words in episodic recognition tasks (Clark, 1992; Glanzer & Adams, 1990, 1990; Hockley, 1994; MacLeod & Kampe, 1996; Malmberg & Murnane, 2002; Reder et al., 2000; Schulman, 1967); however, on the basis of multiple recent findings, it has become clear that the frequency of prior exposure can help encoding. The evidence suggests that highly frequent and familiar items are easier to encode,

[1] For example, a Google Scholar search for the query ["trace strength" AND memory] returns 2520 results

to associate to one another, to bind to an episodic context and to hold in WM (WM; e.g. Diana & Reder, 2006; Reder, Liu, Keinath, & Popov, 2016; Reder et al., 2007).

The interpretation of these findings is not widely accepted, however, and while many researchers recognize that people can retrieve stronger memory traces more quickly and more accurately, few acknowledge that the prior strength of a trace affects the ease with which it can be encoded, associated or manipulated. As a result, few memory models posit a role for prior item strength in episodic learning, knowledge formation, or WM capacity (for a review of WM models, see Oberauer, Farrell, Jarrold, & Lewandowsky, 2016).

Our main goal in this article is to describe a theory of human memory that can account for the multifaceted effects of prior familiarity on memory. In the process, we will (see the Appendix for the organizational structure of the paper):

- introduce a theory of memory, and its mathematical formulation, in which a key novel mechanism is that the encoding, binding and manipulation of stronger items require fewer WM resources (Section "**Overview of the theory**")
- review evidence consistent with the claim that stronger items have an encoding advantage in both long-term memory (LTM) and WM (Section "**Existing challenges for a theory of frequency effects**")
- support a key prediction of the theory, according to which memory for one item is influenced by how many resources are spent in processing preceding items (Section "**Analyses of preceding item strength: novel predictions**")

II. Overview of the theory

The theory is an evolution of the Source of Activation Confusion model (SAC; Reder et al., 2000; Reder & Schunn, 1996; Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997), which itself has roots in the ACT-R cognitive architecture (Anderson et al., 2004). SAC implements a spreading activation theory in which semantic and episodic memory traces are represented as localist nodes in a network. In SAC, memory strength is a continuous value stored within nodes and the links between them; this strength increases through practice and decays with time. The strength of semantic and episodic nodes lead to two different signals, familiarity and recollection, which puts SAC in the class of dual-process models. The model has successfully fit a variety of findings in recognition and cued-recall memory, including the key effects of mirror word frequency (Reder et al., 2000), list length (Cary & Reder, 2003) and list strength effects (Diana & Reder, 2005). For that reason, we imported many of its assumptions.

Despite these successes, the original version of SAC[2], just like other models of memory, did not assume that the current memory trace strength affects the probability of encoding. The lack of such a mechanism prevented it from explaining why stronger items would be easier to encode, bind and manipulate. This core insight was articulated by Reder et al. (2007) and Diana & Reder (2006) on the basis of quasi-experiments that tested memory as a function of normative word frequency, and since then we have accumulated, in support of the theory, a number of critical experimental findings that directly manipulated exposure frequency of the stimuli (Reder et al., 2016; Shen, Popov, Delahay, & Reder, 2018).

The essence of the current theory is that the processing of stimuli, whether for encoding, storage, binding, and retrieval of stimuli, consume WM resources inversely proportional to the strength of the stimulus representation in LTM. This principle can be illustrated by analogy to muscle fatigue. During any physical activity, substrates within muscle fibers fuel contractions of the muscle. These substrates get depleted with every contraction and they recover over time if the muscle is not used. When lifting a heavier weight, muscle fibers use a greater amount of their available substrates. In this analogy, weaker items in memory are like heavier weights – they are more difficult to manipulate, and their retrieval, binding and encoding require more resources.

To be more specific, our theoretical claims are that:

1. The encoding, updating, and binding of stimuli to context and to other stimuli, depend on a limited pool of WM resources
2. These operations deplete the WM resource pool in a continuous manner which in turn replenishes over time[3]
3. These operations consume more resources when the stimuli are weaker, i.e, less familiar[4]
4. As a result of maintaining or manipulating less familiar chunks of information, there will be fewer WM resources available for processing additional stimuli.

---

[2] The original SAC model was published in 1996 and it was applied to cognitive illusions (Reder & Gordon, 1996) and feeling-of-knowing during equation solving (Reder & Schunn, 1996). It was extended to model the mirror-frequency effect in recognition memory in 2000 (Reder et al., 2000). The resource-depletion-and-recovery assumption was added to the model in 2007 (Reder et al., 2007). The current version of SAC revamps many of the underlying equations, but the core mechanisms are shared with the 2007 model. For the rest of the paper, when we refer to "the original SAC model", we refer to the pre-2007 model that lacks the resource-depletion-and-recovery assumption. Whenever we refer just to SAC or "the model" or "the theory", we refer to the current implementation.

[3] This idea bears some similarity to the concept of "ego depletion" (Baumeister, Bratslavsky & Muraven, 2018), according to which volitional acts of self-control and self-regulation draw on a limited resource that recovers over time. Despite the surface similarity, there are distinct differences between our theory and the "ego depletion" concept – first, our conception of resources applies to the storage and maintenance of information, not to executive control, and, second, it is implemented in an explicit computation model

[4] Within this theory, stimuli with stronger memory traces are perceived as more familiar; for simplicity, we will use the terms "stronger" and "more familiar" interchangeably throughout the manuscript (similarly for "weaker" and "less familiar")

We believe that these key new extensions of the theory hold true regardless of how they are implemented. Below we present a complete description of the theory and its computational implementation. Some of the core assumptions are inherited from the previous models of tasks that used SAC (e.g., Reder et al., 2000), and we will note where the new model deviates.

A. Full description of the theory

1. Representation

Information in memory is represented as a set of nodes in an associative network. These nodes reflect semantic or perceptual concepts (e.g., the concept of a "bush" or the representation of a face; "semantic nodes"), events ("I planted a bush yesterday" or "I studied the word "bush" in this experiment"; "episodic nodes"), or contextual information (the internal and external context associated with an experience; "context nodes"). For simplicity, we consider these nodes to be the basic representational units, but in principle they also include detailed underlying semantic, phonemic or lexical features. In other words, the localist nodes are high-level abstractions, and the theory is potentially consistent with different lower level representations. Figure 1 shows a basic schematic of the model representations in a typical paired-associate list learning memory experiment.

We distinguish between stimuli that have existing representations in memory, and stimuli that do not. For example, when people see a novel Chinese character, they have to encode it as a configuration of simpler visual features that they already know, such as a small triangle, the letter X, etc. With repeated experience, the feature nodes get associated to a single node, which eventually, with repetition, becomes a representation of the entire character as a whole. This process, which we refer to as *chunking*, is not unique to our model, and bears similarities to many influential proposals in theories of visual perception (Gilbert, Boucher, & Jemel, 2014; Palmer, 1977), statistical learning (Fiser & Aslin, 2005; Perruchet & Vinter, 1998), and has a long history in memory research (Chase & Simon, 1973; Gobet et al., 2001; Simon, 1974).

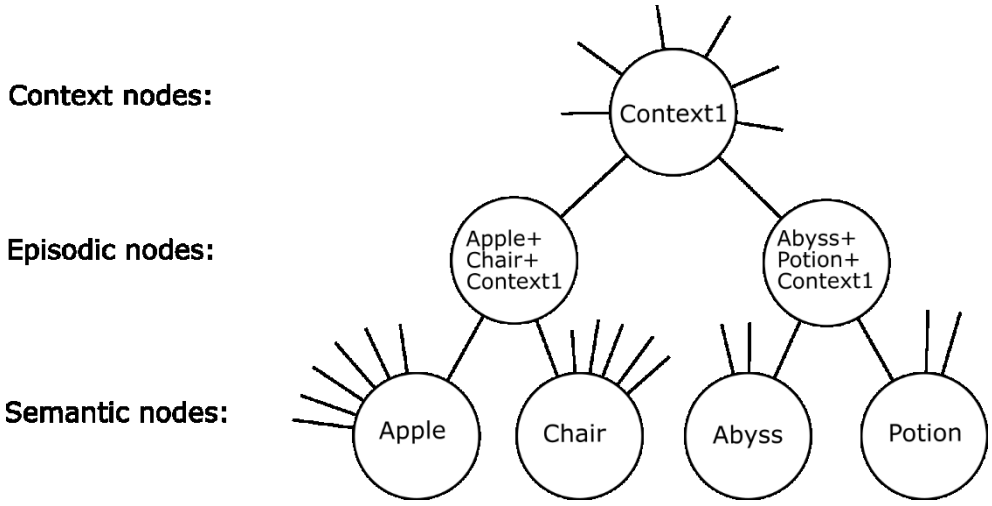

Figure 1. *Illustration of the SAC model structure for a paired-associate experiment. Participants have studied the word pairs Apple-Chair and Abyss-Potion (among others). Each concept has a pre-existing semantic node, which has connections to multiple episodes in which it has been experienced over time. High frequency words (e.g. Apple) are experienced more often, thus have a greater contextual fan of preexisting connections, compared to low frequency words (e.g. Abyss). The current list context also has a separate node, which is connected to all the episodes (i.e. different trials) experienced in the current list. There is a unique episode node that connects all features of an experience, i.e., the two concepts and the context in which they are experienced.*

2. Learning, forgetting, and base-level strength

A long standing distinction in memory research is that between "storage strength/learning" and "retrieval strength/performance" (Bjork & Bjork, 1992; Soderstrom & Bjork, 2015). We take a similar approach in SAC – we distinguish between the base-level strength of a node and its current activation. These values vary as a function of exposure frequency and recency. Base-level strength is determined by the long-term history of exposure to a node, while the current activation represents the availability of the node at the time of test and it depends on what cues are available.

We first describe the properties that characterize base-level strength. Every time a node is activated, its base-level strength increases by an amount specified by Equation 1, and these increments decay as a function of time (specified by Equation 2). These principles are illustrated in Figure 2. Deviating from prior SAC models, in the current model the increase in base-level strength, *s*, depends on how strong the node is at the time of study[5]. Specifically, we now assume

[5] In previous models, each repetition strengthened memory by the same fixed amount, and the summed strength was then log transformed. Log transformation squashes larger values more, which accounts for the diminishing returns of practice. The same principle is behind ACT-R declarative memory module.

that nodes can reach a maximum strength of 1, and that each repetition strengthens the node as a proportion δ (learning rate) of the maximum strength minus its current base-level strength, *B*:

$$s = \Delta B = \delta(1 - B) \quad (1)$$

We initialize new nodes with a base-level strength of δ, because they have no prior strength[6]. As a result of Equation 1, weaker memory traces are strengthened more than stronger traces and the increment size is controlled by the learning rate, δ. For example, a node whose base-level strength is $B = 0.8$ will be strengthened by $s = 0.2 * \delta$, while a node with a base-level strength of $B = 0.1$ will be strengthened by $s = 0.9 * \delta$. Similar equations have been successful in modeling classical conditioning in the influential Rescorla-Wagner theory (Rescorla & Wagner, 1972).

We chose this learning function, rather than the one used previously in SAC and ACT-R, because it has several desirable properties. We wanted the amount of resources depleted from WM to be inversely related to the stimulus current strength, so we needed a way to quantify the resource cost of this operation. Since weaker items are strengthened more (i.e., $\delta(1 - B_{weak}) > \delta(1 - B_{strong})$), if the cost of resources of an increment of size *s* is set to be proportional to *s*, the cost of strengthening weaker items becomes larger.

The base-level strength decays with time, which is one of the main causes of forgetting in the model. The increment in strength from each repetition decays independently[7] of other increments as a function of how much time has elapsed since its occurrence (as illustrated in Figure 2). Thus, at any time, *t*, the base-level strength of a node is:

$$B = B_0 + \sum_{i=1}^{n-1} s_i \times (1 + t - t_i)^{-d}, \quad (2)$$

where s_i is the strength increment produced by the *i*-th repetition, $t - t_i$ is the time since the *i*-th repetition, *d* is the decay rate, and B_0 is the preexisting base-level strength. Due to the power function[8], the initial time value was offset by 1, so that immediately after encoding the increment size is not infinite.

---

[6] Equation 1 specifies the size of the strengthening increment under the assumption that there are sufficient resources available – see Eq. 7a and 7b, which specify how the default strength increment changes if WM resources are insufficient.

[7] see Pavlik & Anderson, 2005, for a neural justification behind this assumption

[8] Researchers have long debated whether forgetting is better described by an exponential or a power function (Rubin & Wenzel, 1996). Consistent with the original SAC model as well as ACT-R, we selected a power decay function; however, SAC has also always postulated an exponential decay for current activation.

Equations 1 and 2 are recursive and in order to know how much each repetition increments the base-level strength, we need to know what the strength was prior to that repetition. Thus, to calculate the strength increment on a certain repetition *n*, we need to combine Equations 1 and 2:

$$s_n = \delta\left(1 - B_0 - \sum_{i=1}^{n-1} s_i \times (1 + t - t_i)^{-d}\right) \quad (3)$$

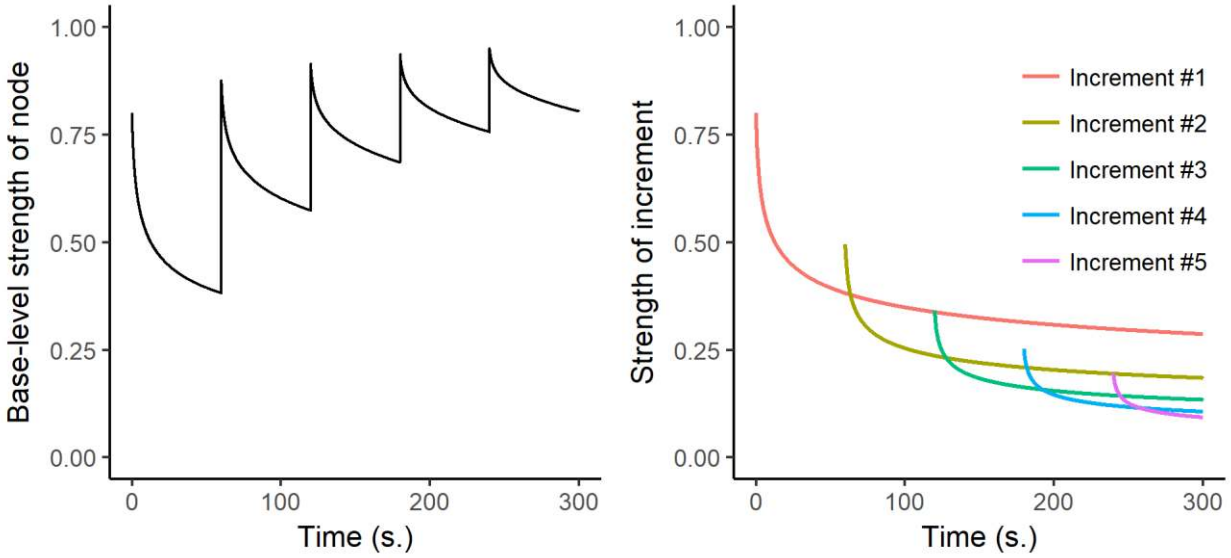

Figure 2. *Illustration of repeated strengthening and decay of a single episode node in the model. The same stimulus is experienced 5 times with 60 seconds in between repetitions. Left panel shows the total base-level strength of the episode node as a function of time. Right panel shows the strength of each individual increment to the node strength. Later increments are smaller, because the size of the increment is a function of the current base-level strength of the node (see the text for details). The simulation for this figure had a learning rate of δ = 0.8 and a forgetting rate d = -0.18 (default values for all simulations).*

The links that connect individual nodes also vary in strength, depending on how often the two nodes have been co-active. The increment and decay of link strength also follow Equations 2 and 3, but with a different decay parameter.

Finally, the prior base level of semantic nodes is a function of word frequency that starts at 0.2 for the lowest frequency items, increases with word frequency and plateaus at 0.4:

$$B_0 = 0.4 - 0.2 \times e^{-0.1\, WFpermillionwords} \quad (4)$$

Word frequency was estimated based on the SUBTLEX norms (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Several factors influenced the design of Equation 4 – 1) the prior base level should vary between 0 and 1 due to Equation 1; 2) the base level should increase with word-frequency; 3) the function connecting word frequency to the base level should have the same shape as the function between the number of study repetitions and the base level (e.g., Equations

1 & 2). Like all other fixed parameters, the lower and upper bound parameters (0.2 and 0.4) were estimated by fitting the model to Reder et al.'s (2000) data (see the Modeling Details section).

3. Strengthening and binding deplete WM resource

The key novel aspect of our theory is that a shared pool of resources fuels these learning processes, and that these processes cost more in resources when applied to items with weaker base-level strength. Every strengthening operation depletes the same WM resource pool. We assume that people have different amounts of WM resources (Daily, Lovett, & Reder, 2001; Lovett, Daily, & Reder, 2000; Lovett, Reder, & Lebiere, 1999), which is denoted by a W_{max} parameter. Every time a person creates a node/link that is strengthened by an amount *s*, s^2 amount of resources is depleted. Under most circumstances, this defaults to Equation 1, squared:

$$W_{default\_cost} = s^2 = \big(\delta(1 - B)\big)^2 \tag{5}$$

Since items with stronger base-level strength, such as high frequency words, are incremented less, their processing depletes fewer resources. We chose the cost of strengthening to be s^2 because the exponent slightly increased the cost difference between small and big increments relative to the overall cost of the operations, which lead to better fits for most models.

We also assume that the resource pool replenishes over time (also see, Buchler, Faunce, Light, Gottfredson, & Reder, 2011; Reder et al., 2007). We assume that the resource pool recovers at a linear rate function of time since the last operation, $t - t_i$, and the remaining resources at time t_i, such that:

$$W_t = \min\big(W_{max},\, W_{t_i} + \, w_r(t - t_i)\big) \tag{6}$$

where, w_r is the recovery rate. WM depletion and recovery are illustrated in Figure 3. The left panel shows the available resources at the beginning and end of each study trial in a single list that contains both weak and strong items (i.e., items whose existing representations differ in strength). Weaker items deplete more resources and the available resources are reduced when more items in a row are weak. The right panel shows the available resources on average after studying strong and weak items in either pure lists of only weak or strong items or mixed lists of both weak and strong items.

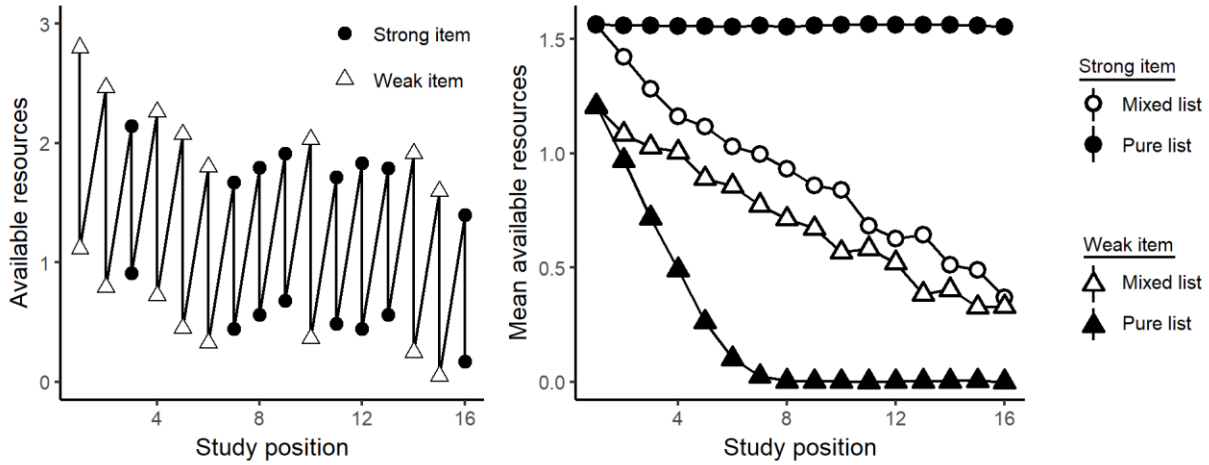

Figure 3. *Illustration of resource depletion and recovery in the model. Left panel – amount of available resources at the beginning and end of each trial during a single study list, as a function of item position on the list and its current strength. The model assumes that resources recover during the inter-stimulus-interval, leading to available resource being higher at the beginning of each trial relative to the ending of the preciding trial. Right panel – the mean available resources at the end of each trial in pure and mixed lists of weak and strong items. Model parameters: learning rate of* $\delta = 0.8$*, resource capacity of* $W_{max} = 2.8$*, and recovery rate* $w_r = 0.45$.

An important issue is what happens when the default cost of a process is more than the remaining resources. One option is that encoding fails because of insufficient resources. However, a simpler modeling alternative is that the system uses whatever resources remain, and the strength increment in Equation 1 and 3 is adjusted by the proportion $\sqrt{\frac{W_t}{W_{defaultcost}}}$:

$$s = \min\left(\sqrt{\frac{W_t}{W_{defaultcost}}}, 1\right) \times s =$$

$$\min\left(\sqrt{\frac{W_t}{{s_n}^2}}, 1\right) \times s =$$

$$\min\left(\sqrt{W_t}, s\right) =$$

$$\min\left(\sqrt{W_t}, \delta(1-B)\right) \qquad (7a)$$

Thus, if the default learning requires more resources than there are available, the strength of the memory trace is incremented by the square root of the remaining resources, $\sqrt{W_t}$ (the square root is due to the square exponent in Eq. 5).

A similar situation arises when multiple stimuli have to be encoded at the same time. For example, if you have to remember a display of several different items, then, depending on their number, there might not be sufficient resources to encode them all. One possibility is that the

system allocates the default proportion of resources to as many items as possible and fails to encode the remaining items. Another possibility is that all k stimuli share the resources proportionally to their default cost, such that the strength of item i is increased by[9]:

$$s_i' = \min\left(\sqrt{W_t \times \frac{(1-B_i)^2}{\sum_{i=1}^{k}(1-B_i)^2}}, \delta(1-B_i)\right), \qquad (7b)$$

where B_i is the base-level activation of item *i.* If the number of items encoded at the same time is small enough, then each node's increment is the default. However, if there are too many items, then they share the resource proportionally to their needs. Equation 7b is a generalized version of Equation 7a, which is a generalized version of Equation 1. When k=1, Equation 7b reduces to Equation 7a; when sufficient resources remain, Equation 7a reduces to Equation 1.

Here we give an example of how the resource depletion dynamics are affected by the sequence of events during an experiment. Compare a trial on which a low-frequency word (LF; 1 occurrence per million) is studied with a trial on which a high-frequency word (HF; 100 occurrences per million) is studied. According to Equation 4, the semantic node's base-level strength is 0.22 for the LF word, and 0.4 for the HF word. Assume that at the beginning of each of these trials, the amount of available WM resources is 0.9. When presented with a word in an experiment, participants first retrieve and strengthen the semantic node for that word, and only then create or strengthen the episode node, and the links between those nodes. From Equation 1, the semantic node for the HF word would be strengthened by 0.8 * (1-0.4) = 0.48, which according to Equation 5 would cost $0.48^2 = 0.23$ resources. The semantic node for the LF word would be strengthened by 0.8 * (1-0.2) = 0.64, which would cost $0.64^2 = 0.41$ resources. Thus, after strengthening the semantic node, the resources remaining would be 0.9-0.23 = 0.67 on the HF trial, and 0.9-0.41= 0.49 on the LF trial. Creating an episode node at full strength in each case would require $0.8^2 = 0.64$ resources. On the HF trial, this is less than the remaining 0.67 resource, so the episode node is created at full strength. On the LF trial, however, only .49 resources remain, and according to Equation 6, the episode node would be created more weakly with a strength $\sqrt{.49} = 0.7$.

4. Interaction between WM load and item strength

Previously we predicted that the difference between strong and weak items would increase as the WM load of the task increases (Reder et al., 2016; Shen et al., 2018). This prediction was based on an intuition and therefore it is important to simulate the interaction effect between word frequency and WM load on the episode node strength. Figure 4 shows the results of this simulation.

[9] Whether the resource is spread amongst all items or whether a fixed amount is allocated for a limited number of items is currently under debate in the literature. See the General Discussion for more on this topic.

As shown in the right panel, when there are sufficient resources (W > 1.05), there is no difference in the episode base-level strength for HF and LF items; however, with an increase in WM load, there is a concomitant decrease in available resources (from 1.05 to 0.41), driving up the difference between HF and LF items to a maximum. At that point there are no longer sufficient resources to build the episode node when encoding LF words (see the preceding paragraph for an example). Decreasing the available WM resources further (from 0.41 to 0.25) reduces the HF-LF difference. Note that the range over which the HF-LF difference increases is four times larger than the range over which it decreases, suggesting that in most cases we should see an increase in the word-frequency effect with higher WM loads, unless the load is so high that few resources remain.

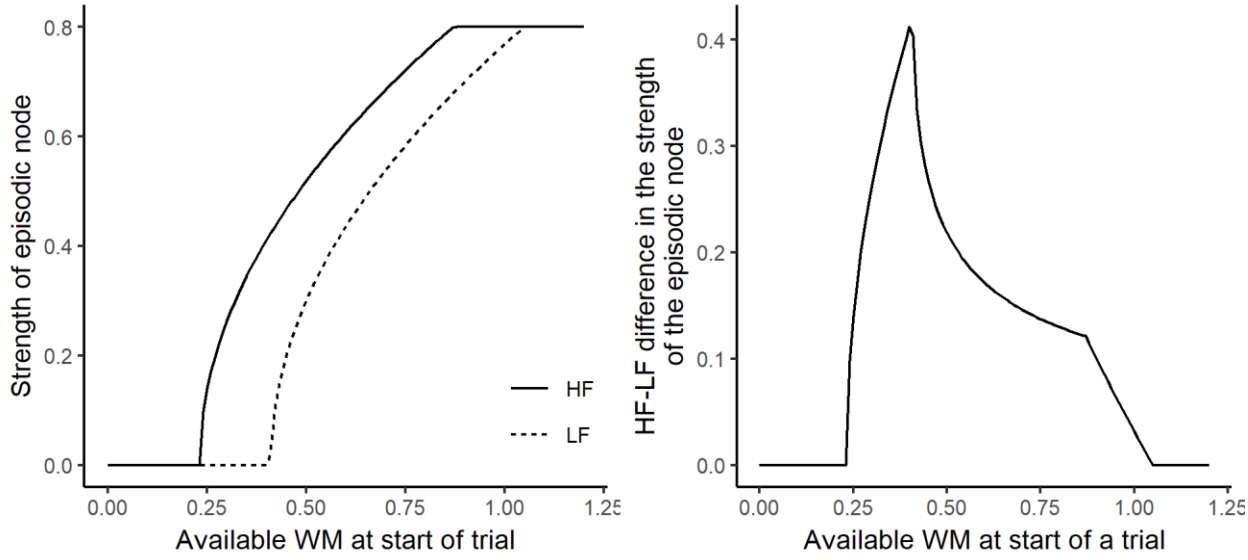

Figure 4. *A simulation of how the difference in the strength changes for the episode nodes encoding HF vs. LF items as a function of WM resource availability (inverse of WM load). The left panel shows the episodic node base-level strength for HF and LF, and the right panel shows the difference between the two. Simulation with a learning rate of δ = 0.8.*

5. Current activation and spreading activation

Nodes are also characterized by their current level of activation. Nodes become active when a concept is perceived, or when activation spreads from other nodes. In contrast to the base-level strength, which decays according to a power-law, the current activation decays exponentially, and it is dependent on the node base-level strength:

$$A = B \times A_{boost} \times e^{-\gamma t} \tag{8}$$

where B is the current base-level strength of the node, t is the time since activation, γ is the exponential decay parameter, and A_{boost} is the amount of activation received by the node.

The *multiplication* of the activation boost by the node base-level strength deviates from earlier versions of SAC, where the activation was *added* to the base-level strength. This formulation leads to more gradual changes in current activation. For example, if the activation boost is added to rather than multiplied by the base-level strength, very weak nodes with strength close to 0 would get a big boost compared to non-existent nodes, which presented a problem for some of our simulations. With the multiplication equation, the current activation gradually increases from 0 without discontinuities.

Under normal conditions, *directly perceived* nodes become maximally active (i.e. A=1), which means that they receive an activation boost of 1/B. The result is that stronger items require less boost to be activated, which could potentially account for perceptual fluency effects (Jacoby, 1983), repetition priming (Huber, Clark, Curran, & Winkielman, 2008), the reduction in fMRI activation for repeatedly presented items (Horner & Henson, 2008), and the fact that HF words elicit less fMRI activation than LF words (Chee, Westphal, Goh, Graham, & Song, 2003).

Nodes can also be activated by their neighbors through spreading activation. When a person perceives the word "dog", this activates not only the semantic node "dog", but also all events in which it was experienced, as well as other concepts related to it. Following Reder et al. (2000) and Anderson, Bothell, Lebiere, & Matessa (1998), we assume that all nodes connected to a source of activation compete with each other – the activation each node receives depends not only on the strength of its connections with the source nodes, but also on the strength and number of links emanating from each source node:

$$A_{boost,r} = \sum_{s=1}^{n} \left(A_s \times \frac{S_{s,r}}{\sum_{i=1}^{k} S_{s,i}} \right) \tag{9}$$

where $A_{boost,r}$ is the boost in activation in the receiving node, A_s is the activation of the source node, $S_{s,r}$is the strength of the link between the source and the receiving node, and $\sum_{i=1}^{k} S_{s,i}$ is the summed strength of all links emanating from the source node. In summary, as the number and strength of outgoing links from a source node increases, the activation that reaches any individual target node decreases. This assumption is based on decades of research on fan effects in memory, which show that it is more difficult to retrieve information when the cue has a greater fan of associates (Anderson & Reder, 1999; Anderson, 1974; Schneider & Anderson, 2012). It has also succeeded in modeling many key results in the literature, such as effects of list length (Cary & Reder, 2003), list strength (Diana & Reder, 2005), and the mirror-frequency effect in recognition (Reder et al., 2000). Equation 9 reflects contextual competition during retrieval and is one of the key mechanisms that counteracts the storage benefits of word frequency, because HF words have greater contextual fan (see Reder et al., 2007). Contextual fan is represented by the $\sum_{i=1}^{k} S_{s,i}$ parameter in Equation 9, and it is a function of SUBTLEX measures of contextual diversity (CD), which correlates positively with word frequency (van Heuven et al., 2014):

$$\sum_{i=1}^{k} S_{s,i} = S_{newlinks} + 0.2 - 0.2 \times e^{-0.1\,CD} \quad (10)$$

where $S_{newlinks}$ is the strength of novel links created during the experiment, and CD is the SUBTLEX contextual diversity measure (percentage of films in which the word appears). Thus, the prior contextual fan varies between 0 and 0.2 (see the Modeling details section for a description of how these parameters were estimated).

6. Memory retrieval

Memory retrieval begins by activating a certain set of cues, depnding on the nature of the retrieval task, which then spread activation to their associated nodes. Figure 5 illustrates how memory retrieval occurs for four different types of memory tasks in which participants study pairs of words. In *item recognition*, the current item and context nodes are activated as cues, and they spread activation to their attached episode nodes. In *associative recognition*, both of the cue words and the context spread activation to attached episode nodes. In *free recall*, only the context node is activated, and thus a relevant episode node receives less activation than it would in the other three tasks. *Cued recall* is similar to item recognition in that the cue word and the context nodes are both activated, and successful retrieval of the target occurs if the episode node is above threshold. Finally, *serial recall* (not illustrated), is achieved via positional representations, such that if participants study a series of items, each item is bound to a node representing its serial position in the list. We presume that since binding requires memory resources, positional representations would be used only when the recall task requires it, as in serial but not free recall. Retrieval in serial recall occurs in a similar fashion to cued recall, but the cues are the positional nodes and the context, rather than the cue items.

Once activation has spread, the outcome of memory retrieval depends on whether the current activation level of the episode node and/or the semantic node, depending on the task, passes a retrieval threshold. SAC is a dual-process model, in which the activation of the episode node represents the recollection signal, while the semantic node's activation level represents the familiarity signal (Reder et al., 2000; Yonelinas, 2002). We follow the signal detection theory tradition and assume that there is noise in the signal and that the probability of a response is the area to the right of a threshold under the normal distribution curve with a mean equal to the node's activation. Thus, if the corresponding node activation is equal to the retrieval threshold, the probability of a response is 50%.

There are differences in retrieval dynamics between recognition and recall tasks. In recognition tests, a "remember" response occurs probabilistically if the *episode* node's activation is above a threshold:

$$P(rem\,|A_{epi}) = P(A_{epi} \geq \theta_{epi}|\,\sigma_{epi}) \quad (11.1)$$

where A_{epi} is the activation of the episode node, θ_{epi} is the episodic retrieval threshold, and σ_{epi} is the standard deviation of the noise added to the episodic node's activation. If the episode node is retrieved successfully, people can recollect the study episode and respond "remember". Even though there are no episodic nodes in memory for new test items, these items can still activate episodic nodes for items with which they share certain feature overlap, leading to spurious recollection due to partial matching (for an in-depth discussion of this issue, see Oates, Reder, Cook, & Faunce, 2015).

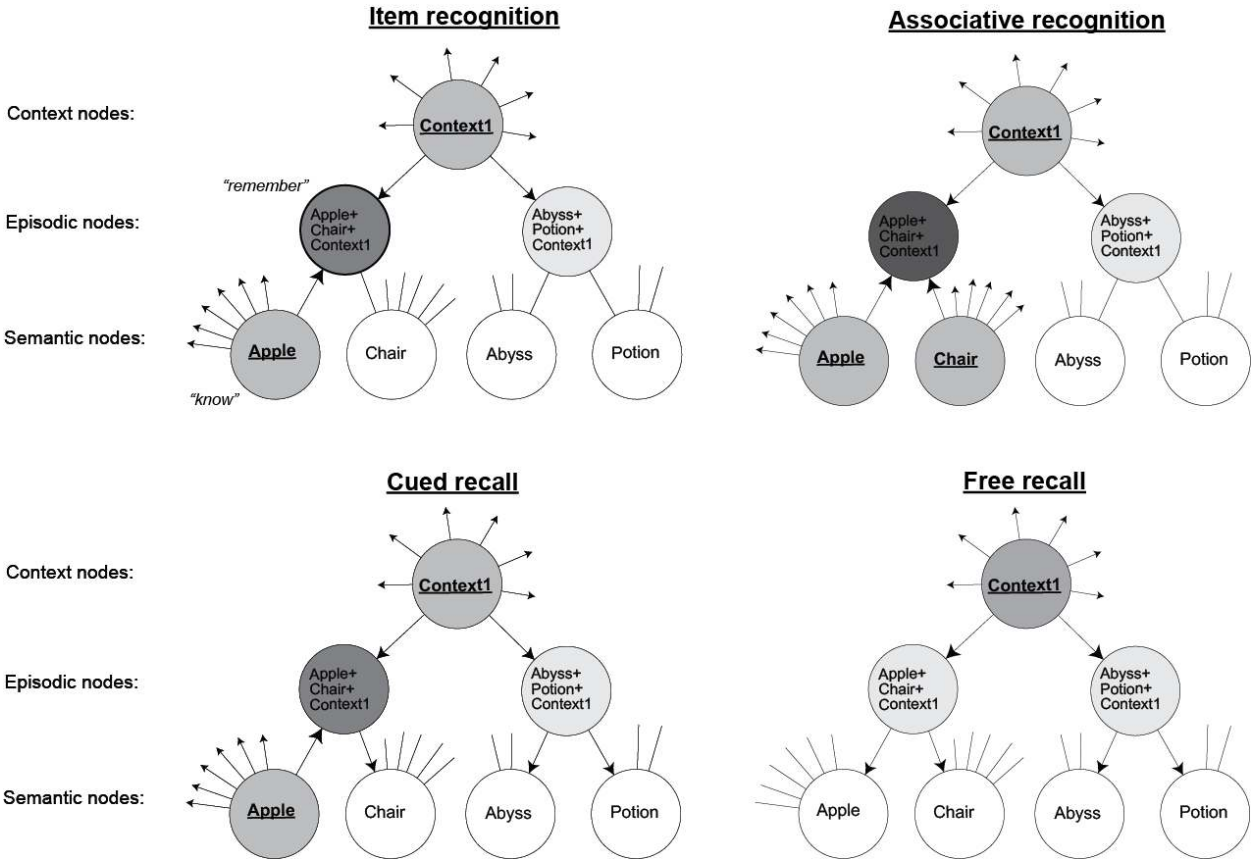

Figure 5. *Illustration of spreading activation for four different test type (see Figure 1 for a description of node types). Participants study pair associates (e.g. apple-chair), and are tested in one of four ways: 1) item recognition (is "apple" old or new?), 2) associative recognition (is "apple-chair" old or recombined?), 3) cued recall (what word was associated with "apple"?) or 4) free recall (recall all words presented in the previous list). While the contents of memory is the same, the amount of activation reaching the episode nodes differs depending on which cues are presented. Underlined text represents the cues for each retrieval task (i.e., context1, apple and/or chair). The shade darkness of nodes represents their activation levels. The episode node activation is highest for associative recognition, medium for item recognition and cued recall, and lowest for free recall. Remember responses and accurate recall probabilities are based on activation of the episode nodes. Know responses and false alarms are based on the semantic node activation*

If the episode activation is below the retrieval threshold, people can still judge an item as old, if it appears familiar. To determine whether an item is sufficiently familiar, the *semantic* node is evaluated instead. We assume that the activation of the semantic node is evaluated before activating it maximally, thus its activation is simply its prior base-level. A "know" response occurs if the *semantic* activation is above a threshold; otherwise, the item is judged as "new":

$$P(know\ |A_{sem}) = P(A_{sem} \geq \theta_{sem}|\ \sigma_{sem}) \times [1 - \mathrm{P}(rem|A_{epi})] \qquad (11.2)$$

In free, cued and serial recall, the semantic node plays no role (except for spreading activation in cued recall), and we assume that a response can be retrieved if the episode node's activation is above its retrieval threshold:

$$P(recall\ |A_{epi}) = P(A_{epi} \geq \theta_{epi}|\ \sigma_{epi}) \qquad (11.3)$$

Figure 6 illustrates how the activation values are transformed to proportion of hits and false alarms for recognition tests, and into recall probabilities for recall tests.

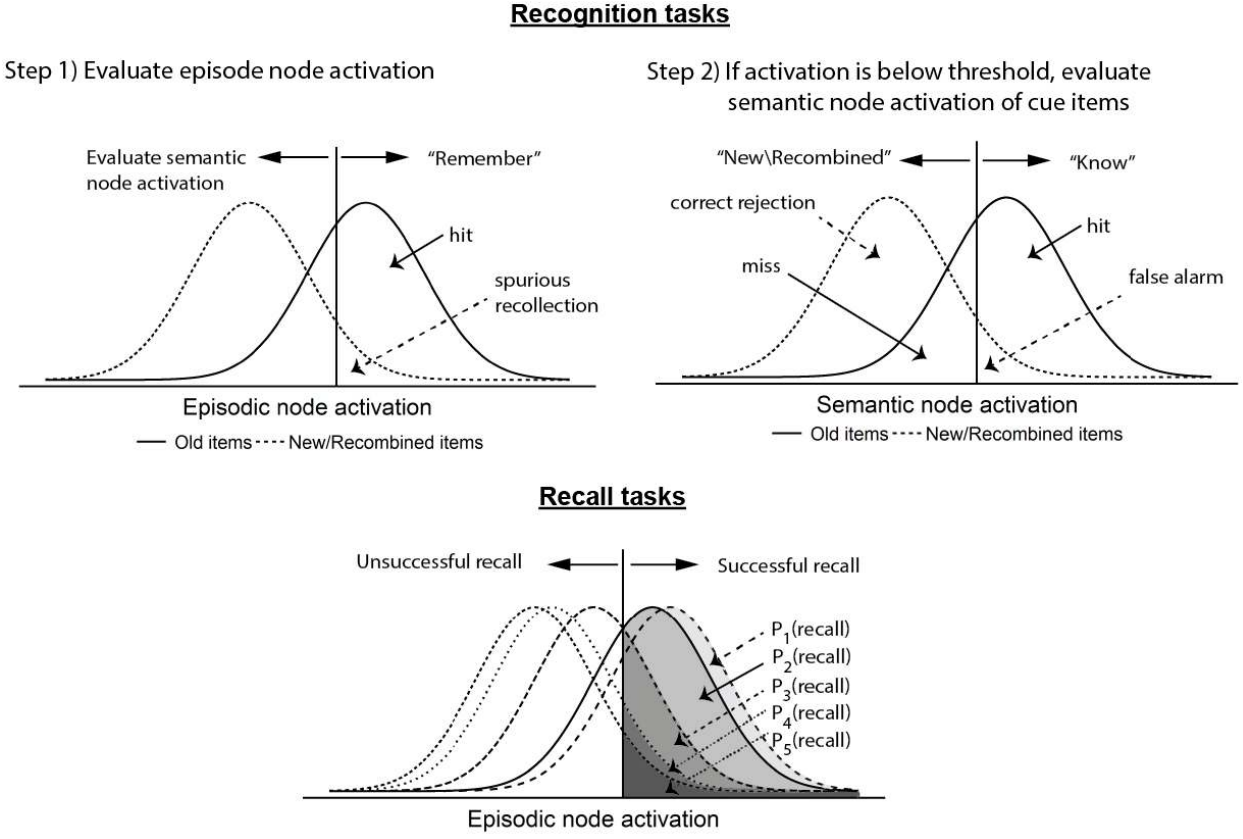

Figure 6. *Illustration of retrieval dynamics in recognition tasks (top) and recall tasks (bottom). Vertical lines represent recognition/retrieval thresholds. Arrows point to the proportion of the activation distribution to the right of the threshold. For recall tasks, the five distributions reflect the activation of five episodic nodes for five different items in memory; P_i(recall) reflects the probability of recalling item i (an arbitrary index).*

7. *Modeling details*

SAC is a process model that takes a sequence of trials that is given to a participant and performs each operation on a trial-by-trial basis. This results in an episodic and semantic activation value for each test trial, which are converted into response probabilities as described in the previous section. The response probabilities were computed through simulations. The number of simulation runs depended on whether we had access to the trial-level data of the study being fit. For studies for which we had access to the trial-level data, we ran one simulation for each participant – this is because given a specific trial sequence the model always generates the same predictions. Thus, the total number of simulations per study depended on the number of participants involved in the study. For studies for which we did not have access to the trial-level data, we generated 100 different trial sequences respecting each study's procedure.

We fit the model by generating predicted response probabilities for each trial and each participant, then summarizing the response probabilities over all subjects and separately for each condition of interest. The parameters of the model were optimized to reduce the root mean squared error (RMSE) between the summarized model predictions and the observed data. The optimization was performed using the downhill simplex algorithm as implemented in Python's Scipy library. We kept most of the parameters fixed across experiments. The decay rates for the base-level strength and current activation of the nodes and the strength of the links were imported from previous versions of SAC (e.g., Reder et al., 2000).

The default delta learning rate and the parameters that convert word frequency into prior base-level strength were estimated anew since the learning function differs from the one implemented in previous model versions. To estimate them, we fit the revised SAC model to Reder et al.'s (2000) data, that involved a continuous item recognition experiment, in which we manipulated the number of repetitions for each word and the words' normative frequency. Figure 7 shows the proportion of Remember and Know responses as a function of presentation number and word frequency, as well as the SAC model fit. The prior base-level strength for the lowest frequency words was estimated to be 0.2, and 0.4 for the highest frequency words. The estimated learning rate was 0.8. These values were used as defaults in all subsequent simulations. Table 1 describes all model parameters and Table 2 summarizes the parameter estimates for the 14 simulations presented in the paper. The modeling code, data and analyses scripts are available at https://github.com/venpopov/prior-item-effects.

Table 1 Description of SAC parameters

Model parameter	**Description**
d_n	Power decay rate for node base-level strength
d_l	Power decay rate for link strength
y	Exponential decay rate for current activation
δ	Learning rate for base-level strength
W	Total WM resource capacity
w_r	WM recovery rate

θ_{epi}	Retrieval threshold for episodic nodes
σ_{epi}	Standard deviation of the noise added to episodic activation
θ_{sem}	Retrieval threshold for semantic nodes
σ_{sem}	Standard deviation of the noise added to semantic activation

Table 2 Parameter estimates for SAC models

#	Study	dn	dl	y	δ	W	w_r	Θ_{epi}	σ_{epi}	Θ_{sem}	σ_{sem}	rmse
1	Reder et al (2000), Exp. 1	-0.18	-0.12	0.2	**0.8**	**3.000**	**0.750**	**0.398**	**0.369**	**0.506**	**0.139**	0.034
2	Clark (1992)	-0.18	-0.12	0.2	0.8	3.000	**0.663**	**0.412**	**0.210**	**0.391**	**0.180**	0.027
3	Malmberg & Nelson (2003), Exp. 2	-0.18	-0.12	0.2	0.8	3.000	**0.864**	**0.101**	**0.263**	-	-	0.014
4	Criss & McClelland (2006)	-0.18	-0.12	0.2	0.8	3.000	**0.995**	**0.100**	**0.332**	-	-	0.006
5	Criss et al (2011), Exp. 1 & 2	-0.18	-0.12	0.2	0.8	3.000	**0.584**	**0.333**	**0.304**	-	-	0.016
6	Criss et al (2011), Exp. 3	-0.18	-0.12	0.2	0.8	3.000	0.584	**0.332**	**0.207**	-	-	0.030
7	Ward et al (2003), Exp. 3	-0.18	-0.12	0.2	0.8	**2.800**	**0.450**	**0.527**	**0.356**	-	-	0.013
8	Miller & Roodenrys (2012)	-0.18	-0.12	0.2	0.8	**3.614**	**0.437**	**0.648**	**0.443**	-	-	0.032
9	Hulme et al (1997), Exp. 2	-0.18	-0.12	0.2	0.8	**5.871**	**0.215**	**0.383**	**0.539**	-	-	0.031
10	Hulme et al (2003), Exp. 1 & 2	-0.18	-0.12	0.2	0.8	**3.837**	**0.330**	**0.572**	**0.469**	-	-	0.026
11	Hulme et al (2003), Exp. 3	-0.18	-0.12	0.2	0.8	**6.832**	**0.827**	**0.463**	**0.390**	-	-	0.036
12	Malmberg & Nelson (2003), Exp. 3	-0.18	-0.12	0.2	0.8	**5.000**	**0.966**	**0.270**	**0.054**	-	-	0.025
13	Diana & Reder (2006), Exp. 1	-0.18	-0.12	0.2	0.8	3.000	**1.114**	**0.091**	**0.020**	-	-	0.035
14	Cox et al (2018)	-0.18	-0.12	0.2	0.8	3.000	**0.970**	**0.143**	**0.067**	-	-	0.024
15	Buchler et al (2008), Exp. 1	-0.18	-0.12	0.2	0.8	3.000	**0.850**	**0.134**	**0.358**	-	-	0.026
16	Aue et al (2017), Exp. 1,2 & 4	-0.18	-0.12	0.2	0.8	3.000	**0.520**	**0.332**	**0.219**	-	-	0.007

Note: **bold-underlined** parameters were free to vary in estimating the model. The remaining parameters were fixed. Dn, dl, and y were imported from the prior version of SAC

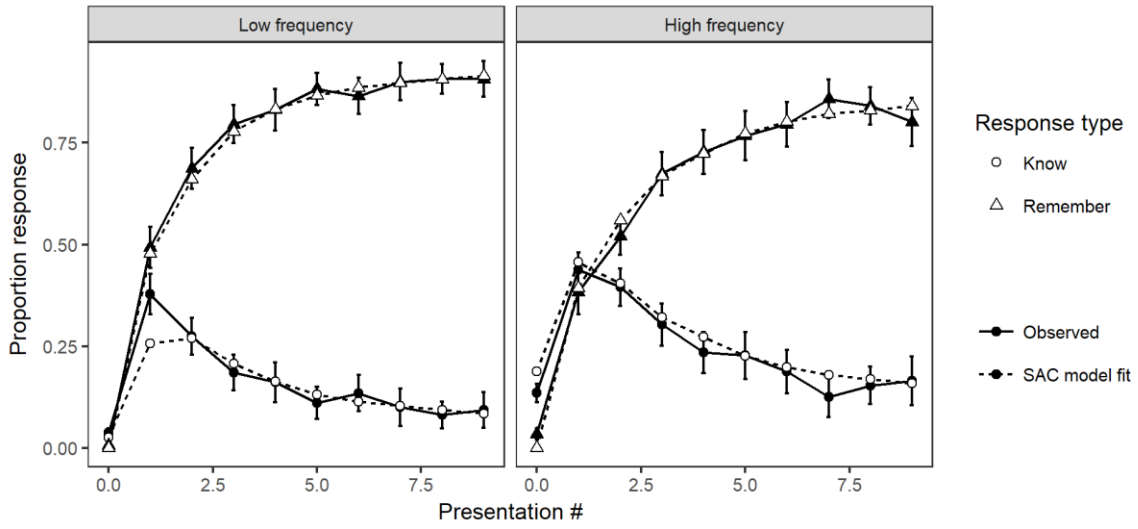

Figure 7. *Proportion of Remember and Know responses in continuous recognition task in Reder et al (2000) for low and high frequency words as a function of repetition number for each repeated word. Open points represent the new fits of the updated SAC model. Error bars represent 95% CI.*

III. Existing challenges for a theory of frequency effects

Rather than reviewing the evidence for all of the theory assumptions, which has been done elsewhere (e.g., Diana et al., 2006; Reder, Park, & Kieffaber, 2009; Reder et al., 2000) we have a more specific focus – we will review evidence consistent with the claim that weaker items deplete more WM resources during memory formation (see also Diana & Reder, 2006; Reder et al., 2007). The argument goes as follows – first, performance on WM tasks is greater for HF items, a fact demonstrated by both quasi-experimental and experimental studies. We argue that while potential confounds with other variables might explain effects in quasi-experiments, the results from Reder et al. (2016), Shen et al. (2018) and the additional analyses of those data finesse any interpretation problems from quasi-experiments **(Section "Challenge 1: Stronger items consume fewer WM resources").** Second, items with stronger memory representations are easier to bind to other items and to an experiential context in order to form novel episodic traces in LTM **(Section "Challenge 2: Item strength facilitates memory formation")**. Finally, other items are also easier to store when presented close in time to stronger items, which suggests that LTM formation and binding might draw on a limited resource that is depleted with each operation and recovers over time **(Section" Challenge 3: Strength of some items affects encoding of other items").** In addition to reviewing well established findings, we present a reanalysis of multiple datasets in which we demonstrate that memory performance for one item depends on the frequency of the items that immediately preceded it during study **(Section "Analyses of preceding item strength: novel predictions")**. Throughout the text we will show and refer to simulation results and fits to existing data. The modeling code, data and analyses scripts are available at https://github.com/venpopov/prior-item-effects. Before we review evidence for each of the claims presented above, two qualifications are in order.

Normative word frequency. One of the main variables which we use as an indication of the existing strength of memory traces is normative word frequency. Normative word frequency is a quasi-experimental variable because one cannot randomly assign words to be of either low or high frequency. As a consequence, there has been a lot of doubt whether normative word frequency has independent effects on memory after controlling for confounded factors such as orthographic and semantic distinctiveness, concreteness, word length, etc. (Cox, Hemmer, Aue, & Criss, 2018; Maddox & Estes, 1997). Our model assumes that HF words have stronger memory traces, because the strength of a memory trace increases with repeated experience. Nevertheless, alternative explanations are always possible, and we will finesse this issue in two ways. First, we discuss at length experiments that make alternative explanations unlikely. Second, we show that similar results hold when frequency is experimentally manipulated by differential training in the lab (Reder, Angstadt, Cary, Erickson, & Ayers, 2002; Reder et al., 2016; Shen et al., 2018; also see Nelson & Shiffrin, 2013).

Difference between number of chunks and chunk strength. A related concern is whether frequency, either natural or experimental, directly affects learning, or whether infrequent items are simply represented as multiple chunks. Numerous studies have shown that it is not the absolute amount of processed information that limits WM capacity, but it is rather the number of chunks

into which this information can be compressed and organized (Miller, 1956; Simon, 1974; for recent reviews, see Cowan, 2001; or Gobet et al., 2001). Researchers may disagree about the specific number of units that constitute the limit of WM capacity (Cowan, 2001; Gobet & Clarkson, 2004; Luck & Vogel, 1997; Miller, 1956) or about the mechanism responsible (Cowan, 2001; Gobet et al., 2001; Oberauer et al., 2016); however, most theories of WM treat chunks as all or nothing. Current verbal learning theories do not see chunks as varying in strength, and, as a consequence, they do not consider that variation in chunk strength would affect WM capacity[10]. Some theorists have argued that exposure frequency in differential training studies does not have direct effects on memory, and that what we call "weak items" might reflect items that do not have a chunked representation. Thus, weak items might often lead to worse memory, because they have to be stored as several chunks[11]. We will show that contrary to this proposal, WM performance is worse with weak items even when there is evidence that they have been chunked.

A. Challenge 1: Stronger items consume fewer WM resources

1. Natural word frequency and WM

Our review begins with tasks which require the temporary storage of a few items into STM. Immediate serial recall is higher for HF words compared to LF words in pure lists (lists containing only HF or LF; Caplan, Madan, & Bedwell, 2015; Hulme et al., 2003; Hulme et al., 1997; Watkins, 1977) and WM change-detection performance is greater for more familiar naturally occurring stimuli (Xie & Zhang, 2016, 2017).

Could these frequency/familiarity effects be explained by some other processing difference between HF and LF words, such as differences in 1) study strategies, 2) semantic associability, 3) speech rate or 4) frequency of rehearsals? The effects occur with both intentional and with incidental encoding, suggesting that it is not due to different study strategies (Morin, Poirier, Fortin, & Hulme, 2006). Immediate serial recall and non-word repetition[12] are also better for non-word syllables that occur more frequently in polysyllabic words (Nimmo & Roodenrys, 2002). Since these are meaningless stimuli, it is unlikely that the effect of frequency is due to greater semantic associability of HF items. Better immediate memory for HF words is also difficult to account for by other variables, such as faster speech rate, because the effect remains even when suppressing articulation (e.g. Tehan & Humphreys, 1988) or when accounting for speech rate (e.g. Hulme et al., 1997; Roodenrys, Hulme, & Brown, 1993). Finally, short-term recall of multi-digit

[10] In contrast, some visual WM models similarly presume that items can be stored with variable precision and more resources can be spent to store visual features with greater precision (for a review, see Ma et al., 2014).

[11] We thank K. Oberauer, E. Awh & N. Cowan for pointing out this possibility to us

[12] In a non-word repetition task participants have to repeat a multisyllabic non-word immediately after the non-word has been presented, which is a measure of STM (Gathercole & Baddeley, 1989).

numbers is higher if a HF pair of items is rehearsed during a delay interval between the encoding and retrieval of the digits (Humphreys et al., 2010). This result shows that holding a pair of LF words in WM hurts the subsequent recall of other items currently maintained in WM, which is incompatible with the alternative explanations discussed above. The fact that frequency effects occur in all of those cases suggests that item strength per se might have a direct effect on the WM consumption. These effects are summarized in Table 3.

2. Experimental familiarization and WM

The studies reviewed above involve stimuli whose familiarity is naturally occurring and it is possible that some other stimulus property is responsible for the results. In two recent studies we gathered direct experimental support for the proposal that weaker chunks deplete more WM resources compared to stronger chunks (Reder et al., 2016; Shen et al., 2018). Both studies involved an extensive training procedure in which we familiarized participants with novel Chinese characters at different exposure frequencies. The familiarization task had hundreds of trials of visual search training over nine to twelve sessions during three to four weeks, three sessions per week. For each participant, sets of visually similar characters were randomly selected to be seen either at high or low frequency during the visual search task (20:1 ratio).

In the first study Reder et al., (2016) measured WM performance for HF and LF characters at the end of training by using an N-back task. Participants responded whether the current stimulus was the same as the one presented N trials previous, where N could be 1, 2 or 3. Participants had to actively maintain the last *n* items in WM, bind each of them to a corresponding serial position, and rapidly update that binding on every trial (Owen, McMillan, Laird, & Bullmore, 2005). As predicted, the N-back performance was better for HF characters, supporting the view that frequency in and of itself leads to differences in WM performance.

In the second study, Shen et al. (2018) provided converging evidence that differential familiarity of stimuli also impacts performance in a different WM task. At the end of training, on each trial of a WM task, Shen and colleagues asked participants to memorize two pairs of character-digit associations, and then to hold them in WM while they solved an algebraic equation. After solving the equation, participants had to identify the characters and recall the corresponding digit associated with each character (e.g. 情 = 3 and 洋 = 7). The characters held in WM were either LF or HF. The equations differed on two dimensions – number of steps (one vs. two) needed to solve for x (e.g., $3x = 6$ vs. $3x - 2 = 7$), and whether the solution required the substitution of constants from the digit span (e.g., 情$x - 2$ = 洋).

Previous research with this task has shown that holding a larger set of numbers in WM hurts equation solution performance, and that this effect increases as the equations become more complex either by increasing the number of steps or by requiring variable substitution (Anderson, Reder, & Lebiere, 1996). If LF characters indeed require more WM resources to be maintained and processed, then solving the equations should be more difficult when people hold pairs of LF characters associated with digits in WM, and this effect should increase with increasing equation complexity. This is exactly what we found. Importantly, it was more difficult to solve the equations

even in the 2-step no-substitution condition, where participants did not have to use the characters they were holding in memory, making it unlikely that the results are due to interference or differential retrieval efficiency from LTM. Furthermore, participants performed better in the recognition and recall task with HF characters.

3. Discounting alternative explanations

Are low-frequency items chunks? A potential criticism is that these results might reflect differences between chunked and non-chunked stimuli, and not differences in chunk strength *per se*. Some have suggested that LF characters were not exposed often enough for people to develop unified representations of each character, and that as a result, multiple features had to be stored independently in WM for LF characters.

While this is possible, at best it is only part of the story for two reasons. First, this explanation cannot account for the quasi-experimental results involving real words. Furthermore, while not included in the Reder et al. (2016) short report, we did directly test whether HF and LF characters reached chunk-like representations. If the alternative explanation described above is correct, we would expect performance in the visual search training task to be negatively correlated with the number of simple features in each character. On the other hand, if these characters become chunked, we should see the effect of number of features diminish with each training session.

We calculated the number of features present in each character based on existing orthographic vector representations (Xing, Shu, & Li, 2004; Yang, McCandliss, Shu, & Zevin, 2009)[13]. These vector representations are based on an orthographic analysis of the characters and prior behavioral work (Xing et al., 2004). We then refit the mixed-effects regression models of Reder et al. (2016) by including the number of features and its interaction with session number and frequency as predictors.

Figure 8 shows that during the first session, reaction times in the visual search task depended on the number of basic features within a character ($\Delta AIC = -424$, $\chi^2(1) = 426$, $p < 0.001$ for the main effect of number of features). Importantly, the effect gets smaller with each training session for both frequency classes ($\Delta AIC = -12$, $\chi^2(1) = 14.60$, $p < 0.001$ for the session by number of features interaction) and while it disappears earlier for HF characters, by Session 9 both HF and LF characters show no effect of the number of features within a character ($\Delta AIC = -41$, $\chi^2(1) = 43$, $p < 0.001$ for the interaction between session, frequency and number of features). Post-hoc comparisons revealed that the slope for number of features is higher for LF characters during sessions one, two and three ($z_{session1} = 4.76$, $z_{session2} = 3.11$, $z_{session3} = 2.33$), but that there was no difference between the slopes for HF and LF characters in the remaining sessions (all z's range from -2.08 to 1.31). By the end of training both HF and LF were chunked successfully.

[13] We thank Xiaonan Liu for pointing us to these representations

We also looked at whether the number of features within a character affected performance in the N-back task. Figure 9 shows that the number of features did not affect performance ($\Delta AIC = 0$, $\chi^2(1) = 2$, $p = .158$ for the main effect of number of features, and $\Delta AIC = 1.3$, $\chi^2(1) = 0.740$, $p = .391$ for the interaction with frequency).

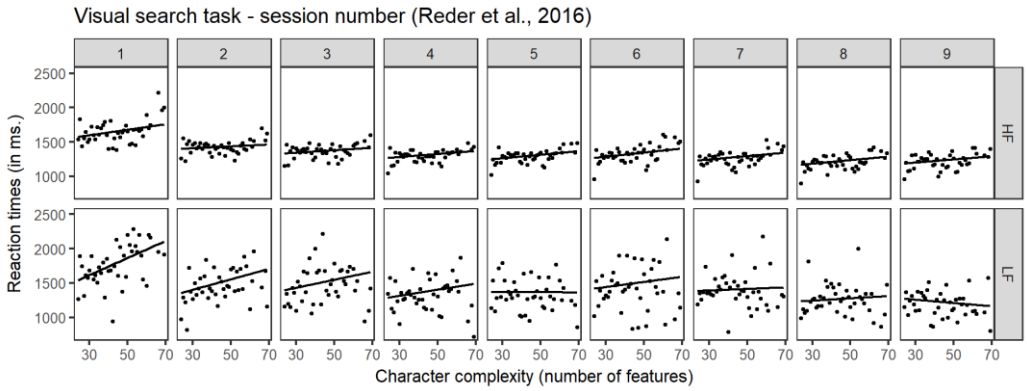

Figure 8. *Reaction times in the visual search task for HF (top row) and LF (bottom row) characters for sessions 1 through 9 (columns), depending on the Chinese character complexity, which was estimated as the number of features in its orthographic vector representations (Xing & Li, 2004; Yang et al., 2009). The data for LF characters is noisier around the slope lines because there are 20 times fewer observations per data point.*

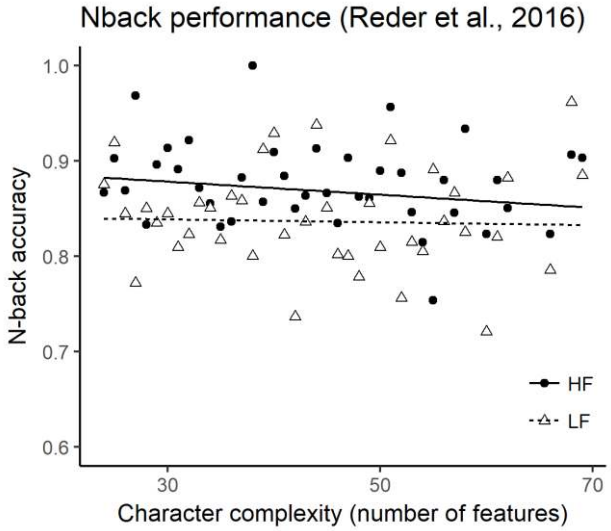

Figure 9. *N-back accuracy depending on the Chinese character complexity, which was estimated as the number of features in its orthographic vector representations (Xing & Li, 2004; Yang et al., 2009)*

Is it just the ease/speed of retrieving representations? A second potential criticism is that the results do not reflect differences in WM consumption, but rather the ease of encoding and/or retrieving the character representations from LTM. The N-back task progresses quickly, and if it

is more difficult to retrieve and store the representation of LF characters this might lead to a decrease in performance that is not due to different WM resource consumption[14].

To test this idea, we performed an additional analysis in which we derived a "processing efficiency" index for each character and each subject individually based on performance in the visual search task. We focused on trials where the target was absent and all distractors were from the same set. On these trials, participants had to do an exhaustive serial search, comparing the target against each distractor, and in the process retrieving the representation of each distractor in turn[15]. On these trials, the reaction time is the sum of retrieval and comparison processes for each distractor. Thus, we used the average reaction time over all trials in which a distractor appeared as a proxy measure for the processing efficiency of that character. For example, if a distractor appeared on 3 trials in which the reaction times were 900, 1000, 1100 ms., then the efficiency index for that character was 1000 ms. The measure was transformed into z-scores and reversed, such that larger values represented more efficient processing.

If our index of processing efficiency reflects differences in character processing, we should expect it to predict performance in the N-back task. Indeed, as can be seen from Figure 10, performance was higher for more efficiently processed characters, and the effect was strongest in the 1-back task, weaker in the 2-back, and non-existent in the 3-back. The important question, however, is whether there remains an effect of character frequency after accounting for the effect of processing efficiency. We repeated the mixed-effects regression analyses reported in Reder et al. (2016) with N-back level, character frequency and processing efficiency as predictors of accuracy. The effect of character frequency remained significant even after accounting for processing efficiency (Figure 11), and, importantly, the interaction between character frequency and WM demands that was originally predicted by Reder et al. (2016) emerged significant.

One limitation of this analysis is that our measure of processing efficiency may not be completely valid or reliable. Some unexplained variance in the N-back task might still be due to differences in processing efficiency between HF and LF characters. Nevertheless, this analysis still decreases the likelihood that processing efficiency is a sufficient explanation. The fact that at the 1-back level the frequency effect could be entirely attributed to processing efficiency differences, suggests that the measure is valid enough to explain performance at low WM loads; at the same time the same measure is not sufficient to explain all the variance at higher WM loads, suggesting that the frequency effect at the 2-back and 3-back levels might indeed be due differences in node strength rather than processing efficiency. The processing efficiency explanation also cannot account for other converging evidence, such as list-composition effects, presented in the remainder of the paper.

---

[14] We thank K. Oberauer for suggesting this alternative explanation

[15] This assumption is based not on theories of visual search, but on theories of perceptual expertise, which presume that when familiar visual stimuli are encountered, the cognitive system automatically and efficiently retrieves their representations from LTM (Logan, 1988; Palmeri, 1997; Palmeri, Wong, & Gauthier, 2004; Palmeri & Tarr, 2007).

The correction for encoding efficiency was necessary because characters differed naturally in complexity. We have since replicated the results of Reder et al. (2016) with better controlled stimuli, but identical training procedure. We used artificial animals called Fribbles, which are three-dimensional combinations of various features and are available online courtesy of M. Tarr (Tarr, 2018). After the visual search training, N-back performance was lowest for low-frequency Fribbles, middle for high-frequency Fribbles, and highest for pictures of familiar animals. Importantly, as shown on Figure 12, the difference between the three familiarity conditions increased with N-back level, supporting the hypothesis that familiarity of stimuli interacts with WM resources.

4. Summary

The findings presented in Reder et al. (2016), Shen et al. (2018), the Fribbles replication of Reder et al. (2016), and the additional analyses presented here provide converging evidence for the claim that HF items require fewer WM resources for processing. The current analyses demonstrated that lower performance in the N-back WM task was not due to failure to chunk the low frequency characters nor due to differences in the ease of encoding and retrieving the characters.

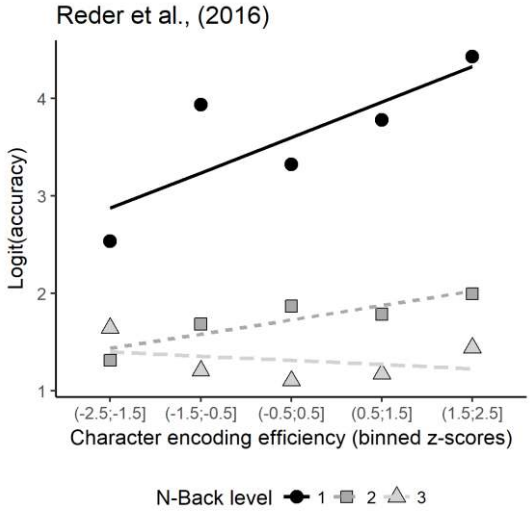

Figure 10. Accuracy (in logit units) in the N-back task depending on N-back level and processing efficiency of each character. Efficiency was estimated based on performance in the visual search task.

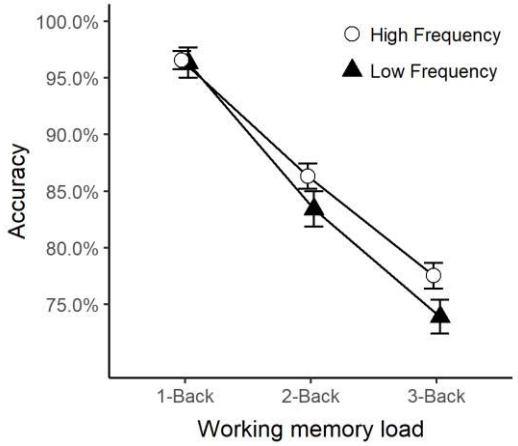

Figure 11. Effect of character frequency on n-back performance after accounting for processing efficiency of each character. Error bars represent 95% CI.

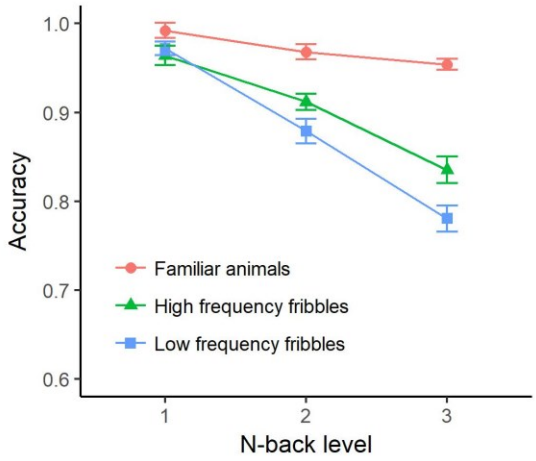

Figure 12. N-back accuracy depending on stimulus type – familiar animals or Fribbles familiarized in preceding visual search task at either high or low frequency. Error bars represent 95% CI

Table 3 Findings consistent with the claim that stronger items consume fewer WM

Serial recall

- Serial recall of HF items better than LF in pure lists (Caplan et al., 2015; Hulme et al., 2003; Morin et al., 2006).
- The effect occurs with both intentional and incidental encoding (Morin et al., 2006).
- Immediate serial recall and nonword repetition are better for nonword syllables that occur frequently in polysyllabic words (Nimmo & Roodenrys, 2002).
- The effect remains even when suppressing articulation (e.g. Tehan & Humphreys, 1988) or when accounting for speech rate differences (e.g. Hulme et al., 1997; Roodenrys et al., 1993).
- Short-term recall of multi-digit numbers is higher if HF words are rehearsed during a delay interval between the encoding and retrieval of the digits (Humphreys et al., 2010).

N-back task

- N-back performance is better when the stimuli are Chinese characters *experimentally familiarized* at high frequency rather than low frequency (Reder et al., 2016).
- N-back advantage for HF characters remains even after accounting for encoding/processing efficiency.
- N-back advantage for HF characters increases with the difficulty of the task (N-back level).
- N-back performance highest for images of familiar animals, medium for images of HF experimentally familiarized novel animals (Fribbles), lowest for images of LF experimentally familiarized Fribbles.

Solving algebraic equations and memory span

- Better performance (accuracy and RT) in solving algebraic equations in task requiring participants to concurrently hold character-digit bindings in WM when those characters were experimentally familiarized at high rather than low frequency (Shen et al., 2018).
- The algebraic performance advantage for HF characters increases as the complexity of the equation increases (Shen et al., 2018).
- Higher subsequent recognition of the characters and associated digits when the characters were HF rather than LF (Shen et al., 2018).

B. Challenge 2: Item strength facilitates memory formation

The main claim we make in this section is that forming and storing novel memory traces is easier for stronger items. By memory formation, we refer to the process that binds the elements of an experience into a unified representation or a memory trace of a single episode. The evidence reviewed below supports our claim in three major ways – items with stronger existing memory traces 1) are easier to bind to an experiential context, 2) are easier to bind to one another, and 3) make it easier to form traces for other items experienced closely in time. These three classes of findings encompass the full spectrum of item strength effects at encoding (see Table 5, Table 6 and Table 7 for summaries), and while various explanations have been offered to account for some of these findings, our theory provides a single mechanism for all.

1. Effects of word frequency on free recall

In free recall, performance is typically better for HF words rather than LF words in pure lists – study lists that contain only one type of frequency items (Balota & Neely, 1980; Deese, 1960; DeLosh & Mcdaniel, 1996; Gillund & Shiffrin, 1984; Gregg, Montgomery, & Castaño, 1980; Sumby, 1963; Ward, Woodward, Stevens, & Stinson, 2003; Watkins, LeCompte, & Kim, 2000). The higher performance of HF words is seen both in terms of overall recall probability, as well as a faster learning rate when studying to a criterion (Sumby, 1963). In pure lists, the effect is parametric – recall is a monotonic function of word frequency (Deese, 1960; although mixed lists might show a U-pattern, Lohnas & Kahana, 2013).

Several explanations have been offered to explain these word-frequency effects. First, HF words are more likely to have preexisting semantic associations to one another, which can facilitate the formation of associations among HF words on the study list (e.g. Deese, 1960; Ozubko & Joordens, 2007; Sumby, 1963). Second, high-frequency words are considered to be more accessible and easier to generate, which facilitates their retrieval during free recall (e.g. Criss, Aue, & Smith, 2011; Madan, Glaholt, & Caplan, 2010). Third, because HF words are more accessible, participants might be rehearsing them more often during study compared to LF words, which strengthens their memory traces to a greater degree (Tan & Ward, 2000; Ward, Woodward, Stevens, & Stinson, 2003). Our model suggests another explanation – forming episodic memory traces is easier for HF words because processing stronger (more familiar) words consumes fewer WM resources, so more words can be so stored.

These mechanisms are not mutually exclusive and may all contribute to the HF advantage in free recall. Yet, the existing explanations listed above cannot account for the full pattern of word frequency effects. HF recall is higher even when accounting for the number and recency of rehearsals (e.g. Tan & Ward, 2000; Ward et al., 2003) or for the degree of the prior semantic associations among HF words. For example, Tan & Ward (2000) used an overt-rehearsal procedure, and showed that in pure lists HF items are rehearsed more often and later in the list. However, a significant effect of word frequency remains even after equating for the number and

recency of rehearsals (Ward et al., 2003). Finally, the proposal that HF words are easier to retrieve cannot explain why the effects disappear in mixed lists (e.g. Ozubko & Joordens, 2007).

Our proposal is consistent with the findings discussed above. For example, if there are more resources available when studying HF words, people might form stronger associations between words studied closely in time, which has been found (Tulving & Patkau, 1962; Ward et al., 2003). These results suggest that part of the HF recall advantage comes not only from encoding individual items, but also from forming more and stronger bindings among adjacent items or items and their temporal context.

One curious free recall result that our model can also explain is that when study items are separated by long intervals filled with a different task, the HF free recall advantage is removed (Gregg, Montgomery, & Castaño, 1980). Gregg et al. (1980), asked participants to study lists of 12 HF or LF words in two different conditions. In the control condition words were presented at a rate of 2 seconds with no pauses in between; in the continual distraction condition, participants had to count down by 3s from a three digit number for 10s after every study item. Thus, in the continual distractor condition, WM resources have 10s to recover until the next study item appears. The authors found not only that overall memory decreased in the continual distraction condition, but that there was no longer a free recall advantage for HF words. We fit our model to Gregg et al.'s data by estimating a single WM recovery rate for all conditions[16]. The empirical results and the model fits are presented on Figure 13.

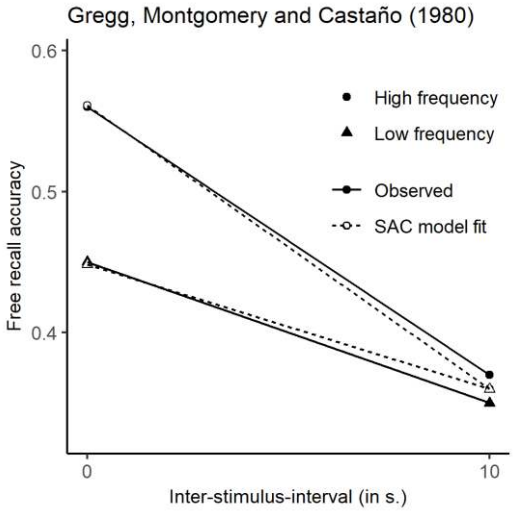

Figure 13. Free recall accuracy as a function of word frequency and inter-stimulus-interval. Empirical data from Greg, Montgomery and Castano (1980) and the corresponding SAC model fit.

[16] For simplicity, our model assumes that backwards counting does not deplete resources, and that resources recover over the distractor period. While we cannot claim with certainty that backwards counting does not require any resources, the model which implements this assumption fit the data well. It is possible that if it does require resources, it depletes them at a negligible rate because participants only have to retrieve the numbers, but they do not attempt to strengthen their representations for future use.

2. Effects of word frequency on item recognition

In contrast to serial and free recall, the most common finding in item recognition memory is that LF words are recognized better, showing more hits and fewer false alarms than HF words (Clark, 1992; Glanzer & Adams, 1990, 1990; Hockley, 1994; MacLeod & Kampe, 1996; Malmberg & Murnane, 2002; Schulman, 1967). This finding, known as the mirror frequency effect, was initially quite challenging for the models of the time, but most current models can explain it (Glanzer & Adams, 1990; Hintzman, 1994; Joordens & Hockley, 2000; McClelland & Chappell, 1998; Reder et al., 2000; Shiffrin & Steyvers, 1997). It might seem that the LF item recognition advantage is inconsistent with the proposal for an HF encoding advantage; however, this is not case, as we explain below.

SAC offers a dual-process explanation for the mirror frequency effect (see Diana et al., 2006; Reder et al., 2000, 2007). In essence, HF words show more false alarms because they have stronger concept nodes, which leads to a greater sense of familiarity for unstudied items. At the same time, HF words are experienced in a greater variety of contexts and situations which makes it much more difficult to retrieve any specific one of them (Equations 9 & 10). This context competition leads to the lower hit rate portion of the mirror frequency effect. What is important for the current argument is that the contextual competition during recognition masks the fact that HF words are stored more easily. This masking does not occur in free recall, because the contextual fan of the words plays no role – in free recall the only cue is the encoding context. Even though the encoding advantage of HF words is often out-weighed by its contextual fan disadvantage in recognition tests, there are a number of cases in which the demands during encoding are great enough to reverse that trade-off (Reder et al., 2007).

In item recognition, the encoding *disadvantage* of LF words is evident with extremely rare words (e.g. Mandler, Goodman, & Wilkes-Gibbs, 1982; Schulman, 1976; Wixted, 1992; Zechmeister, Curt, & Sebastian, 1978). When a wide range of frequency is considered, recognition hits have a U-shape, where very rare words are recognized worse than HF words (Mandler et al., 1982; Zechmeister et al., 1978). Relatedly, recognition increases as a monotonic function of subjective familiarity of rare words (less than 1 in 4 million; Schulman, 1976). These findings appear to be in conflict with results by Lohnas & Kahana (2013) who found that recognition is a monotonic decreasing function of word frequency[17]. However, Lohnas & Kahana (2013) tested

[17] In contrast to Lohnas & Kahana (2013), Hemmer & Criss (2013) found a U-shaped pattern for word frequency on item recognition performance. One difficulty with comparing these two studies is that participants in the Lohnas and Kahana (2013) experiment studied individual words, while participants in Hemmer and Criss (2013) studied word pairs, despite being tested on single item recognition. As we note in our discussion of associative recognition, HF words usually have an advantage in associative recognition because the storage demands are higher compared to a task in which one studies individual items. This causes the storage advantage of HF words to overcome their retrieval contextual competition disadvantage. Similarly, even though Hemmer and Criss (2013) tested item recognition, not associative recognition, the fact that participants studied word pairs might explain the reversal of the recognition frequency effect at the highest frequencies.

words with frequency of more than 2 counts per million, and their lowest bin of stimuli had an average frequency of 21 counts per million. Thus, the recognition disadvantage for very rare words seems to be robust.

One criticism of studies using very rare words is that their recognition might be worse not because they are less frequent per se, but rather because they are unknown to participants (Shiffrin & Steyvers, 1997). This potential criticism is similar to the one we addressed concerning WM in the section "Are low frequency items chunks?" Without a preexisting representation in LTM, the task of remembering a word becomes more akin to associative learning. Associative learning is more difficult than item learning because it requires binding not only the item and the current context, but also the constituent parts of the unfamiliar item. Yet, similar U-shaped results have been found in training studies, in which participants were exposed to pseudowords (Reder et al., 2002) or to three-digit numbers (Maddox & Estes, 1997) at different experimental frequencies prior to a recognition memory task. In summary, when frequency of exposure is low enough, either quasi-experimentally or experimentally manipulated, its encoding disadvantage appears even in item recognition.

Finally, our theory predicts that the hit rate advantage of LF words in recognition should be lower or eliminated when there is a reduced encoding and binding ability, which will overcome their recognition benefit. One group that consistently shows reduced WM capacity and episodic memory binding ability is older adults. As predicted by our account, the typical LF hit rate advantage gets monotonically smaller as age increases, while the recognition of HF words does not change (Balota, Burgess, Cortese, & Adams, 2002). While this effect might be attributed to increased familiarity of LF words with age, other attentional factors also eliminate or reverse the LF hit rate advantage – for example, reducing the study time (Criss & McClelland, 2006; Malmberg & Nelson, 2003) or dividing attention during study (Diana & Reder, 2006)[18].

The next modeling challenge we tackled was the result that reducing study time reduces the hit rate advantage for LF words (Malmberg & Nelson, 2003) or even reverses it (Criss & McClelland, 2006). We chose these studies because timing is critical when we consider that WM resources recover over time. We fit our model to Malmberg and Nelson's (2003) findings by estimating a single WM recovery rate for all conditions (similarly for Criss & McClelland, 2006). Figure 14 shows the empirical data and the fit of the SAC model. The excellent fit of the model reinforces the idea that the reduction of the LF hit rate advantage could be due to insufficient resources available as a result of the fast presentation rate. It is noteworthy that Criss and McClelland's (2006) fastest presentation rate (0.15 s) was faster than that of Malmberg and Nelson (2003) and produced a recognition benefit for HF words in both the model and the experiment. This is because, at very fast presentation speeds, resources recover minimally, which hurts the

[18] One could question whether the LF advantage disappears because of floor effects in all of these studies; however, when we consider the false alarm levels, recognition performance in all of those studies was well above chance levels, making this explanation unlikely (for example, FA rates in Malmberg & Nelson (2003) were 0.37 for HF words and 0.21 for LF words, while hits rates were above .50 in all conditions).

storage of LF words more than the storage of HF words, counteracting the LF recognition advantage due to contextual competition.

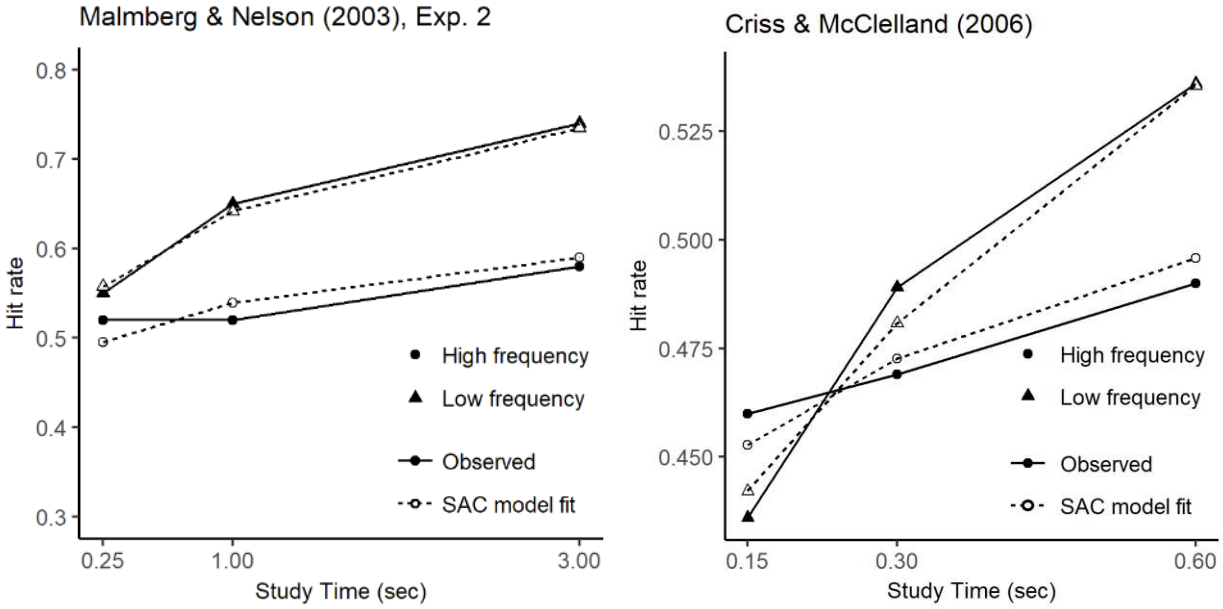

Figure 14. Hit rate for item recognition of low and high frequency words depending on the trial duration during study. Data from Malmberg & Nelson (2003, Exp. 2; left), Criss & McClelland (2006; right) and SAC model fits.

3. Effects of word frequency on associative recognition

The other case in which we observe beneficial effects of HF words on recognition is with associative recognition wherein people study pairs of words and have to discriminate between intact (old words studied in the same pair), recombined (old words studied in different pairs) and novel pairs (new words). Associative learning is more demanding because it requires the encoding of each item, as well as binding the two items together; in our view, each of these operations draws on the same limited resource pool. Consistent with this idea, multiple experiments have demonstrated that pairs (Chalmers & Humphreys, 2003; Clark, 1992, Experiment 1; Clark & Shiffrin, 1992) or triplets (Clark, 1992, Experiment 2) of HF words are easier to recognize than pairs or triplets of LF words (although see Hockley, 1994 for a null effect), even if item recognition for the same stimuli is lower for the HF words (Clark, 1992). Figure 15 shows Clark's 1992 results and the fit of our computational model in which encoding LF words left fewer resources for encoding the association between them.

Our account is also supported by the boundary conditions of frequency effects in associative recognition. First, the HF advantage in associative recognition is present only when the words in the pair are not strongly semantically associated (Martin, 1964). When a strong prior association exists, participants only have to encode the episodic trace and do not have to bind the items to one another, presumably because they are already chunked as an association. As a result, there are sufficient resources to encode both LF and HF word pairs. The HF advantage also disappears with

incidental encoding (Humphreys et al., 2010). With incidental encoding instructions participants do not necessarily attempt to bind the two words together, and the demand on WM is smaller, which results in equivalent performance between HF and LF word pairs. Finally, like the familiarization studies we discussed earlier (e.g. Maddox & Estes, 1997; Reder et al., 2002), repeated exposure to LF words *prior* to the experiment facilitates subsequent associative learning (Chalmers & Humphreys, 2003), demonstrating a *direct* benefit of exposure frequency for association formation.

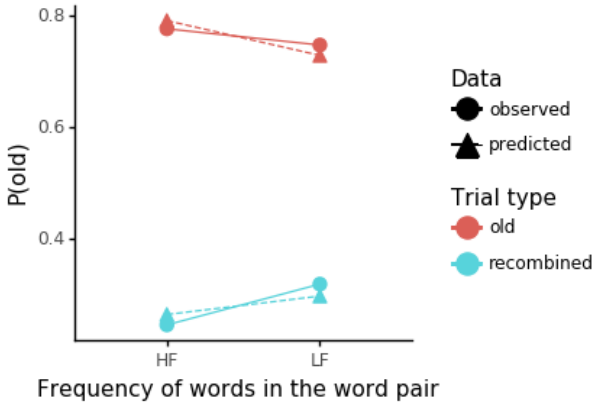

Figure 15. Proportion of old decisions in associative recognition of word pairs containing HF of LF words in Clark (1992), and the fit of the computation SAC model to the data.

4. Effects of word frequency on cued-recall and source memory

The storage and association formation advantage of HF items is also demonstrated in studies involving source memory and cued-recall. In source memory tasks, participants have to retrieve the context in which they experienced a certain stimulus. Results show that is easier to store the study context for items that have preexisting representations (e.g., better recognition when the scene in is reinstated for famous but not non-famous faces, Reder et al., 2013), for items rated as more familiar (DeWitt et al., 2012), and for HF compared to LF words in pure (Popov, So, & Reder, 2019), but not in mixed lists (Osth, Fox, McKague, Heathcote, & Dennis, 2018). According to our theory, the processing of an unknown face (Reder et al., 2013), a LF word (Popov et al., 2019) or a less familiar item (DeWitt et al., 2012) consumes the available WM resources to build a long-term representation, so there are fewer resources available for forming a link between the item representation and the context in which they appear. Furthermore, consistent with the resource depletion explanation, Popov et al. (2019) found that slowing down the presentation rate from 500ms to 750ms to 1000ms linearly decreases the difference in source memory between LF and HF trials. Interestingly, the effect reversed in mixed-lists, and LF words had a source memory advantage in the slowest presentation time (also see Osth et al., 2018), possibly due to the fact that they have fewer pre-experimental associations, leading to less contextual competition (Equations 9 & 10; Reder et al., 2007).

Two comprehensive studies on how word frequency affects cued-recall provide additional support for our position (Criss et al., 2011; Madan et al., 2010). Both studies evaluated separately

the effect of cue and target frequency by manipulating them orthogonally. HF targets were better recalled than LF targets. When it comes to the cue, however, Madan et al. (2010) found no effect of word frequency, while Criss et al. (2011) found that while HF cues were more effective than LF cues, the effect was much smaller than that for the target. Criss et al. (2011) found those results surprising and argued that most memory models would predict a benefit for LF cues, because, for example, they have fewer pre-experimental associations (Equations 9 & 10; Dennis & Humphreys, 2001; Reder et al., 2000). Our explanation for the weaker, but positive effect of cue frequency is that since LF cues take more resources to encode, there are fewer left for forming the cue-target association. That is, as we discussed before in the item recognition section (see also Reder et al., 2007), the trade-off between lower contextual competition for LF words and their encoding disadvantage can mask the positive effects of the former, and the negative effects of the latter. Consistent with these claims, Criss et al. (2011) found in Exp. 3 that when the difference in context diversity of the cues is increased, the effect of the cue frequency is removed or reversed.

We modeled these data by fitting a single recovery rate to both experiments, and by estimating contextual fan from the SUBTLEX contextual diversity ratings for each word in the experiment. While Criss et al. (2011) claimed to have equated context diversity in Experiment 3, the SUBTLEX context diversity rating revealed that this was not the case (see Table 4). A 2x2 ANOVA revealed that HF cues had more context diversity compared to LF cues in both the Low and the High context diversity conditions ($F(1, 261) = 72.87$, $p < .001$). The difference in context diversity between HF and LF cues was bigger in the high context diversity condition ($F(1, 261) = 12.68$, $p < .001$). Since the difference in context diversity between HF and LF cues is bigger in the High context diversity condition, Equations 9 and 10 would predict that the beneficial effect of HF cues would be removed or even reversed, which is what Criss et al. (2011) found. Figure 16 shows the fits of the SAC model to the data from Criss et al. (2011), implementing these assumptions. The model captures the reduced benefit of HF cues compared to HF targets, and also the fact that the cue benefit disappears with high contextual fan, because bigger difference in contextual fan between HF and LF cues in that condition.

Table 4. Mean SUBTLEX measure of cue context diversity in Experiment 3 of Criss et al (2011), depending on the cue frequency and Criss et al's context diversity category. CD = Context diversity

Criss et all CD category	Cue frequency	Cue CD (Mean ± SE)
Low context diversity	HF	8.41 ± 1.32
Low context diversity	LF	1.25 ± 0.17
High context diversity	HF	19.5 ± 2.38
High context diversity	LF	2.61 ± 0.29

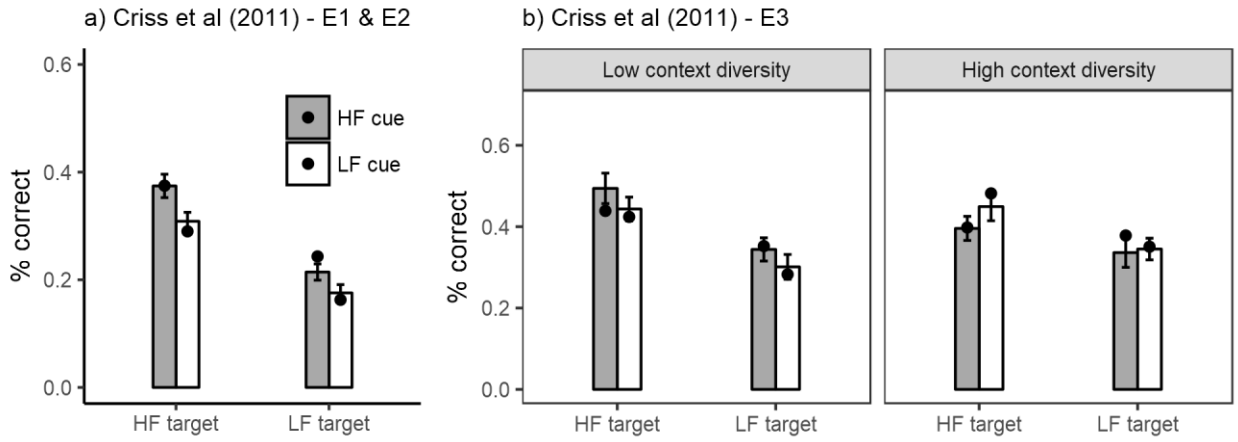

Figure 16. *Cued recall probability as a function of word frequency of the cue and the target in Criss et al (2011, Exp. 1, 2 and 3), and fits of the SAC model (dots) – a) Exp. 1 and 2 combined; b) cued recall depending on whether the words had low or high context diversity (Exp. 3). Error bars represent 95% CI*

5. Experimental familiarization and memory formation

Both Madan et al. (2010) and Criss et al. (2011) depended on normative word frequency. If the explanation based on the contextual fan/frequency trade-off seems convoluted, consider the large-scale familiarization study with Chinese characters that we discussed earlier (Reder et al., 2016). In that study, in which there was no difference in contextual diversity between HF and LF characters, we also showed that cued-recall is better with novel cues that were trained at high rather than low-frequency, contrary to the findings of Criss et al. (2011) and Madan et al. (2010). The cued-recall task was performed at the beginning of each week starting on Week 2. During the study session of the cued-recall task, participants saw two Chinese characters from the same frequency class paired together with an English word. Each character was present in two different triplets on a given list (e.g. the character A was studied in the triples A-B-CHAIR and A-C-TREE), so that participants were forced to attend to both characters to be able to respond with the correct English word. Each week, the study pairings were novel combinations of characters with a previously unpresented English word.

Two results from this study are consistent with the idea that item frequency benefits association formation. First, performance in the cued recall task increased each week, even though the triplet combinations were novel on each list. Importantly, novel pairs of HF characters were better cues for recalling the English words than novel pairs of LF characters and the advantage was as large in the delayed test after 2-4 weeks post-training. Given that characters were randomly assigned to frequency class, the results rule out possible confounding factors for the conclusion that frequency of exposure directly facilitates association formation. Finally, we argued earlier that cue frequency had a small or null effect in Criss et al. (2011) and Madan et al. (2010) studies because of a trade-off of higher contextual competition with less WM demands. The results of Reder et al. (2016) support that explanation: When there are no differences in contextual competition, HF cues will still have a large, positive effect on cued-recall.

Table 5 Findings consistent with the claim that stronger items are easier to encode and bind

Free recall

- Single item free recall of HF is better than LF in pure lists (DeLosh & McDaniel, 1996; Sumby, 1963; Ward et al., 2003).
- Learning rate of HF is also faster, such that with repeated trials, HF word lists reach criterion more quickly (Sumby, 1963).
- Greater temporal clustering of HF words in free recall (Tulving & Patkau, 1962; Ward et al., 2003)
- Dividing attention during encoding eliminates the HF advantage in free recall (Gregg, Montgomery, & Castaño, 1980).
- The frequency advantage in free recall / serial recall is difficult to dilute even with repeated practice. People learned sequences of HF words more easily even after completing 8 trials of repeated free-recall testing on the same words (Sumby, 1963).

Item recognition

- Single item recognition of very low LF is worse than HF (with natural frequency, Schulman, 1976; with pseudo-word familiarization, Reder et al., 2002).
- LF recognition advantage is reduced in older adults (Balota et al., 2002).
- LF recognition advantage eliminated with divided attention during encoding (Diana & Reder, 2006).
- LF item recognition advantage is eliminated with short study durations (Malmberg & Nelson, 2003).

Pair recognition

- Pair associate recognition of HF words is better than LF words (Clark, 1992; Clark & Shiffrin, 1992; Chalmers & Humphreys, 2003; although, Hockley, 1994, finds no effect).
- HF pairs recognized better only when the words are not strongly associated (Martin, 1964).
- The benefit for HF word pairs is removed with incidental encoding (Humphreys et al., 2010).
- Preexposure to LF words improves subsequent associative learning (Chalmers & Humphreys, 2003).

Cued recall

- Cued recall is better for high-frequency cues and targets (Criss et al., 2011; Madan et al., 2010), but only with pure lists (Clark & Burchett, 1994).
- Chinese characters ***experimentally familiarized*** at high frequency are easier to associate to one another as the compound cue and to the English language response term than those familiarized at low frequency (Reder et al., 2016).

Source memory

- Better source memory for more familiar items, but the effect disappears when attention is divided during item encoding (DeWitt et al., 2012).
- Better source memory for HF words in pure lists, but the effect is reduced with longer study times (Popov et al., 2019).

C. Challenge 3: Strength of some items affects encoding of other items (list composition effects)

In this section we examine evidence for another aspect of our thesis – if encoding depends on a limited resource, and the amount of resources spent is an inverse function of the strength of the item being processed, then storing weaker items should lead to lower memory performance for subsequently processed items, because fewer resources remain. We consider mainly list-composition effects, both at the global level (i.e., overall performance differences for pure and mixed frequency study lists) and at the local level (i.e., performance for specific items depending on the strength of immediately preceding or concurrently studied items). See Table 6 for a summary of these findings.

1. Pure vs mixed list paradoxes

A major variable that interacts with word frequency on memory performance is list composition. *Pure study lists* are those that contain only HF or only LF items, while *mixed study lists* are lists that contain both LF and HF items in an either random or alternating manner.

The HF advantage in free recall and source memory tasks holds only for pure lists, and is absent or reduced when stimuli of different frequency are mixed within a list, a phenomenon often referred to as the mixed list paradox (e.g. Gillund & Shiffrin, 1984; MacLeod & Kampe, 1996; Ozubko & Joordens, 2007; Popov et al., 2019; Ward et al., 2003; Watkins et al., 2000). The typical result is that the presence of LF items on the list hurts memory for HF items, and the presence of HF items on the list helps memory for LF items (i.e., performance for HF pure > HF mixed ~ LF mixed > LF pure; see Figure 17). SAC suggests that fewer resources are depleted while storing HF items, which leaves more resources available for storing LF items in mixed rather than pure LF lists. We fit the SAC model to the mixed-vs-pure data of Ward et al. (2003, Exp. 3), whose results can be seen in Figure 17. Estimating a single recovery rate, threshold and noise parameter for all four conditions provided a good fit to the data (eight data points were fit; see Figure 25 for the remaining four).

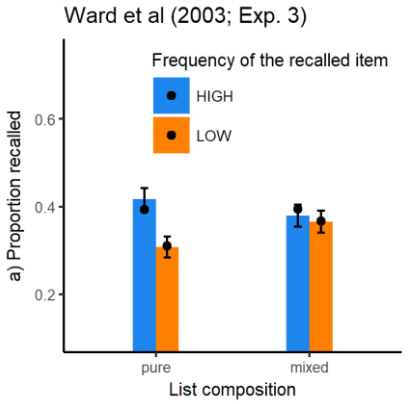

Figure 17. Free recall proportion for high and low frequency words depending on whether they were studied in pure or mixed frequency lists. Data from Ward et al (2003, Exp. 3) and fits of the SAC model (dots). Error bars represent 95% CI

Can the model account for more intricate aspects of list-composition effects? For example, increasing the proportion of LF items on the list monotonically decreases overall performance. Recall accuracy is highest in lists composed of 100% HF items, medium in lists composed of 75% HF items, and lowest in lists composed of 25% HF items (DeLosh & McDaniel, 1996). Similar results hold with word recognition, where the discrimination between LF targets and foils improves as the proportion of HF words on the list increases (Malmberg & Murnane, 2002), and with source memory, where the ability to recall the spatial position of LF cues improves as the proportion of HF words on the list increases (Popov et al., 2019). In SAC, the reduced probability of storing LF items on mixed lists is due to the fact that storing HF words depletes fewer resources, and as a result, the greater the proportion of HF words on the list, the more resources remain to process the LF words.

SAC also predicts that the *precise order* of LF and HF items on a list will impact performance. Indeed, performance in immediate serial recall is better when the first half of each mixed list contains HF words and the second half contains LF words, rather than the reverse; performance on both types of mixed lists is intermediate between pure HF lists and pure LF lists (e.g., HH > HL > LH > LL, where the two letters reflect word frequency within the first and the second half of the lists; see Figure 18, Miller & Roodenrys, 2012; Watkins, 1977). In addition, the word frequency effect increases with serial position in immediate serial recall (Hulme et al., 1997).

One can wonder why order matters when the total amount of resources expended would seem to be the same regardless of order. It would not matter if resources were not limited from below or above; however, resources cannot be depleted below 0 and they cannot recover above the maximum capacity. That means that overall depletion depends on item order. The key question is how much of the resources are available at the time of encoding: When LF words are encountered first, they deplete more resources, leaving less for encoding the second half of the list; the reverse is true when the HF items are encountered first (see Figures 18 and 19 for the fit of the model to Miller & Roodenrys's data, 2012). Relatedly frequency effects in pure lists increase with serial position because additional items compound the effect of frequency in terms of spending more and more resources, leading to an interaction with serial position. Our simulations supported this explanation and the fits of the model to Hulme et al.'s (1997) results are shown in Figure 20.

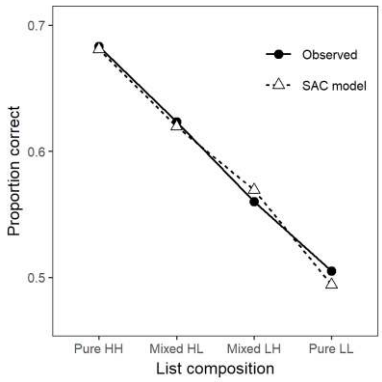

Figure 18. WF effects in pure and mixed lists in immediate serial recall. There are two types of mixed lists – HL are lists in which the first three items were HF, and the last three were LF; LH are lists in which the first three items were LF, and the last three were HF. Data from Miller and Roodendrys (2012) and SAC model fit.

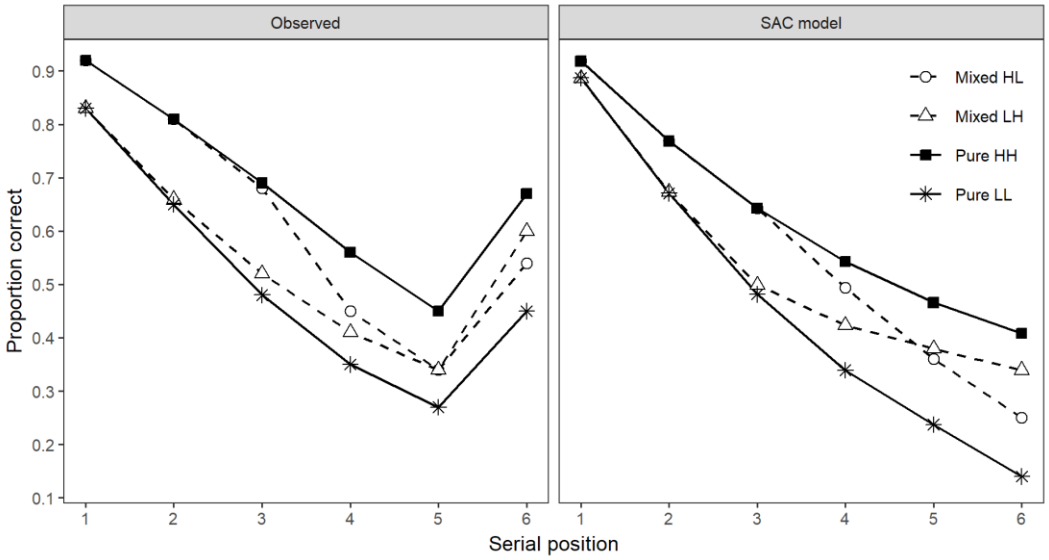

Figure 19. WF effects in pure and mixed lists in immediate serial recall as a function of serial position. There are two types of mixed lists – HL are lists in which the first three items were HF, and the last three were LF; LH are lists in which the first three items were LF, and the last three were HF. Data from Miller and Roodendrys (2012) and SAC model fit. See footnote 21 concerning the recency effect for the final item

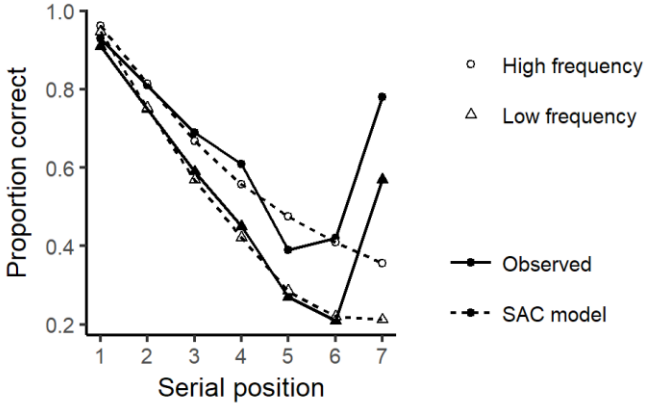

Figure 20. Word frequency interacts with serial position during immediate serial recall. Data from Hulme et al (1997) and fits of the SAC model.

A more in-depth exploration of the serial position curves for pure and alternating lists in Hulme et al (2003) provides further support for the resource-depletion account (see Figure 21). If we focus on the first item in each sequence, we can see that recall for the HF item in lists that begin with an HF item (HFLF lists) is equivalent to the recall of the first item in the pure HF lists, while recall for the LF item in lists that begin with an LF item (LFHF lists) is equivalent to that of the first item in pure LF lists. This makes sense, because resources have been depleted to the same degree on the first trial for pure HF lists and mixed HFLF lists; likewise for pure LF lists and LFHF lists. However, the situation changes for HF items on the second trial. In the second serial position, performance is lower in the LFHF list than in the pure HF list, even though both test an HF word (vice versa for LF items on HFLF lists in the second position). The results are similar when the

lists contain words and non-words instead (Figure 22). The model simulations, which assume that each item is bound to a serial position cue (Anderson et al., 1998; Figures 21 and 22) show the same pattern in the alternating lists, supporting our explanation (see the Model limitations section concerning the absence of a recency effect in the model). In summary, the complex interaction of frequency and serial position in alternating lists supports the resource-depletion account.

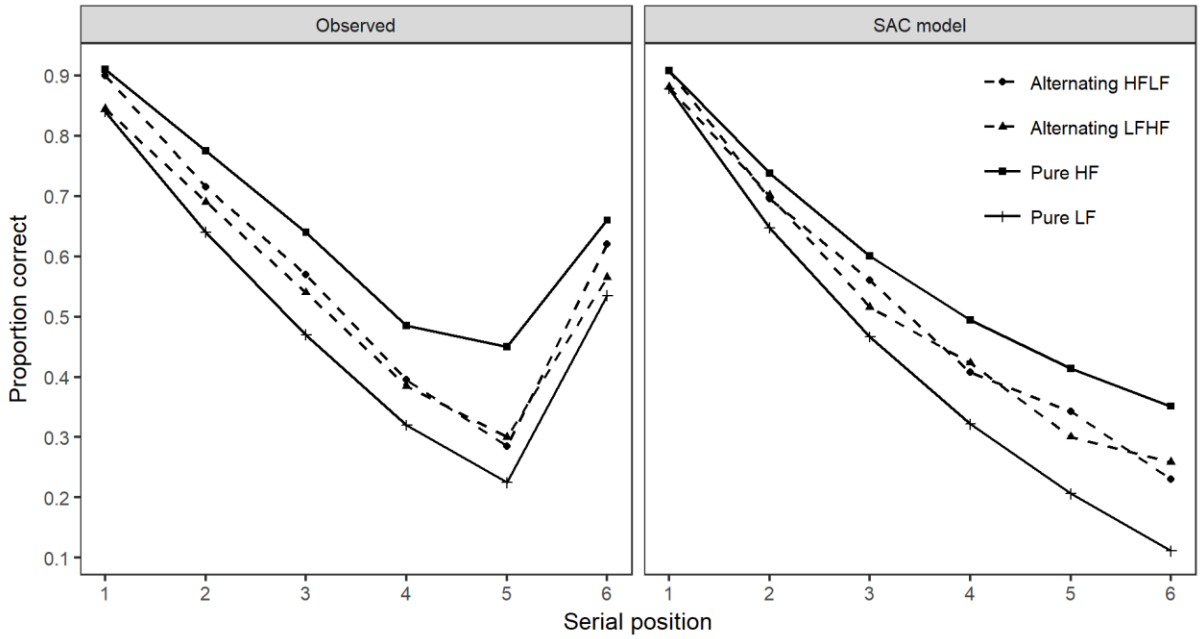

Figure 21. WF effects in pure and alternating lists in immediate serial recall. Data from Hulme (2003; Averaged from Exp. 1 and Exp. 2) and SAC model fit.

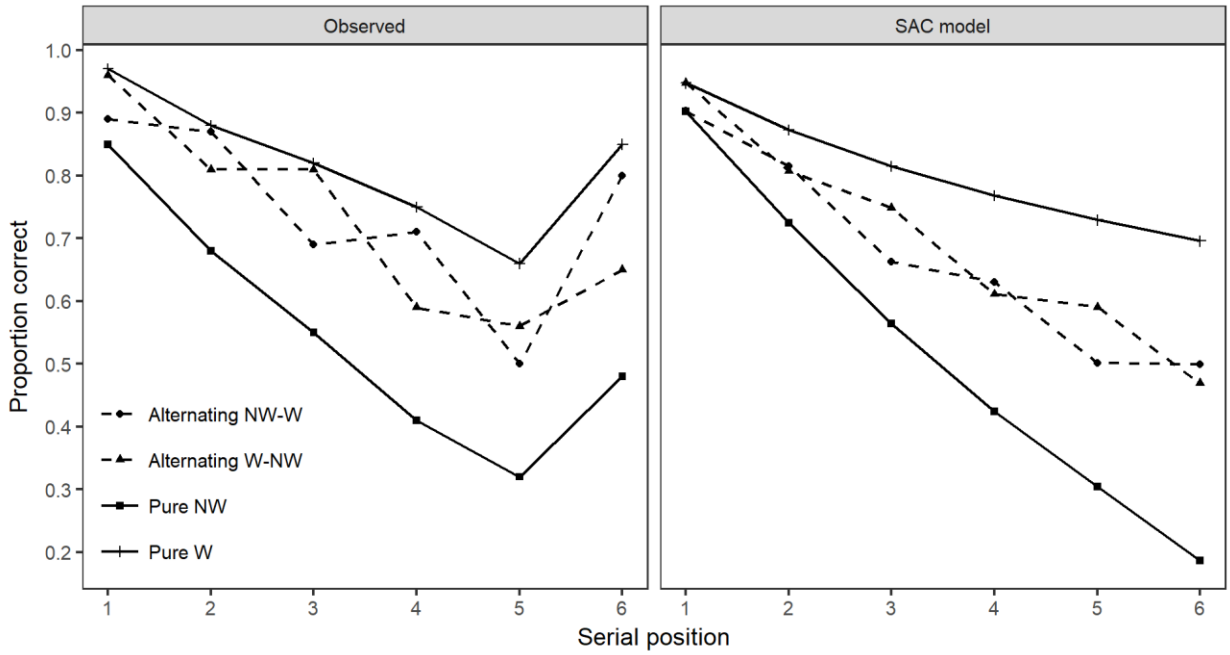

Figure 22. Word vs non-word immediate serial recall in pure and alternating lists. Data from Hulme (2003; Exp 3) and SAC model fit.

2. Frequency of concurrently studied items

A direct test of the idea that LF items require more resources, leaving fewer resources for processing concurrent items, comes from experiments in which participants studied pairs of items, but received an item recognition test for individual items within each pair (Diana & Reder, 2006; Malmberg & Nelson, 2003).

Malmberg & Nelson (2003) asked participants to study pairs of two HF words, two LF words or mixed pairs of one HF and one LF word, and then asked participants to make an *item recognition* decision for only one of the words from each pair. When it comes to the frequency of the tested word, the authors found the typical LF recognition advantage. In contrast, when the other word in the pair was LF rather than HF, hit rates for the tested word were lower, given that the study time was kept relatively short (1.2 sec). This effect is consistent with the idea that LF words require more resources, a conclusion also supported by the fact that LF targets were hurt more than HF targets in the presence of another LF word. Interestingly, the effect of the partner word's frequency disappeared with longer study times (4.0 sec; see Figure 23). Malmberg and Nelson's (2003) interpretation was that LF words require more resources only during the initial encoding stages. While their interpretation is possible, SAC postulates that resources recover over time; as a result, the 4 second condition may provide sufficient time for resources to recover so as to more easily store both items. While a slower presentation rate removes the effect of the frequency of the partner word, the recognition advantage for LF targets remains, because it is due to effects at retrieval, not encoding, specifically the lower contextual fan for LF words makes it is easier to get activation back to the episode node. In contrast, in the 1.2 second condition resources do not recover fully from one trial to the next, impacting LF words more than HF words. To test this explanation, we fit the SAC model to the data by estimating only one recovery rate for all conditions (Figure 23).

One might argue that it is unnecessary to invoke the idea of WM resources to explain these results: it is possible that in Malmberg & Nelson's (2003) study people were merely spending more time encoding LF words, leaving them less time to encode the other word in the pair (e.g., Rao & Proctor, 1984). Diana & Reder (2006) discounted this explanation. In addition to replicating Malmberg & Nelson's (2003) results with word pairs, in a different experiment they asked participants to view pictures of objects with words superimposed on them. Participants had to name the word as quickly as possible but attend to both the word and the picture because their memory would be tested for both. Recognition memory for the pictures, which was tested independently of the words, was better for pictures studied with superimposed HF words. Importantly, there was no significant difference in naming times between LF and HF words (630 vs 607 ms.). Since people allocate the same amount of time to pronounce LF and HF words, and the next trial began immediately after naming the word, it is not likely that these results can be explained by differences in study time, *per se*.

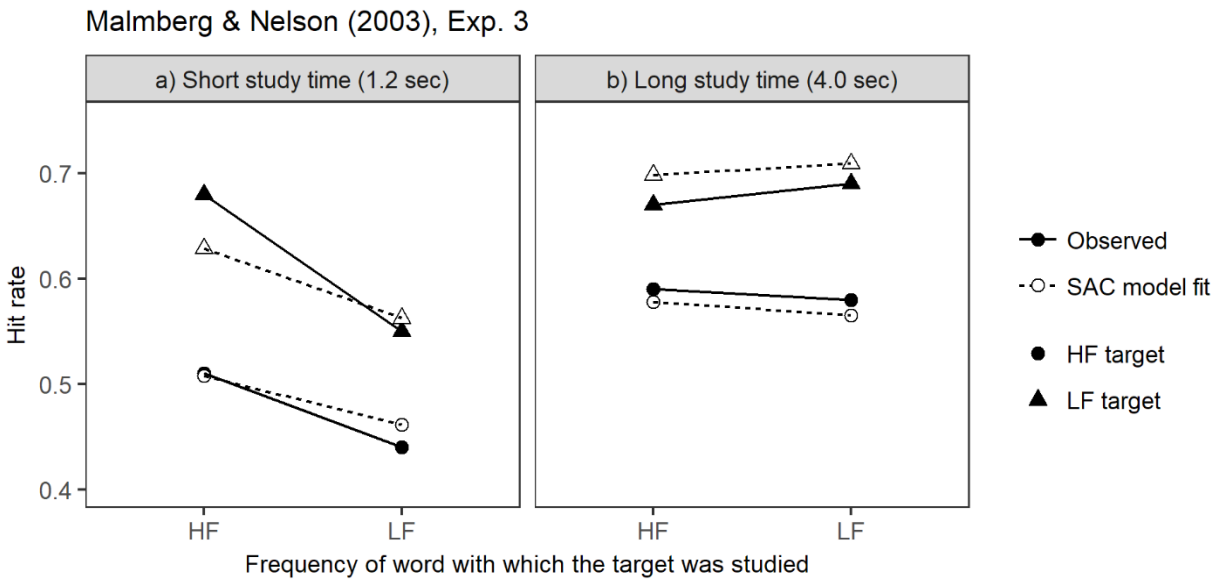

Figure 23. Hit rate for low and high frequency targets, depending on the frequency of the other word in the study pair and the study duration. Data from Malmberg & Nelson (2003, Exp. 3) and the SAC model fits.

Table 6 Findings consistent with the claim that strength of some items affects the encoding of other items

Pure vs mixed list paradoxes

- Free recall advantage for HF words only on pure lists, but not on mixed lists (MacLeod & Kampe, 1996; Watkins, LeCompte, & Kim, 2000; Ward et al., 2003; see Ozubko & Joordnes for a meta-analysis)
- Free recall of HF items improves as the proportion of HF items on the list increases (DeLosh & Mcdaniel, 1996).
- Single item recognition of LF improves as the proportion of HF words increases (Malmberg & Murnane, 2002).
- Serial recall of high-frequency words is worse, while low-frequency words is better, when embedded in mixed lists rather than pure lists (Watkins, 1977; Hulme et al., 2003; Caplan et al., 2015).
- Greater serial recall span when the first half of the list is HF and the second half LF, compared to the reverse (Watkins, 1977; Miller & Roodendrys, 2012)
- the word frequency effect increases with serial position in immediate serial recall (Hulme, 2003; Hulme et al., 1997).

Frequency of concurrently studied items

- Single item recognition is better when the item was studied concurrently with HF rather than with LF item (Diana & Reder, 2006), but only with short study times (Malmberg & Nelson, 2003)

Strength of immediately preceding items

- Preceding item frequency effects on learning of current item (Section IV)

IV. Analyses of preceding item strength: novel predictions

The viability of a model should be judged not only by its ability to fit existing data, but also by its ability to make novel predictions. If SAC's unified explanation for list-composition effects is true, then we should expect that memory should be affected not only by the frequency of the concurrently studied items (Diana & Reder, 2006; Malmberg & Nelson, 2003), but also by the frequency of the immediately preceding items during study. Since LF items deplete more resources, and these resources recover gradually over time, then memory should be worse for items that are preceded by LF items during study. In the remainder of the paper we will present *re-analyses* of multiple existing datasets from different labs that use a variety of memory paradigms, all of which support this prediction.

SAC makes several distinct predictions concerning the effect of the preceding items. Consider Figure 24, which represents a sequence of study items. We expect that memory for item X_k will be a function of how much of the resource was spent in memorizing the immediately preceding items X_{k-1}, X_{k-2}, X_{k-i} etc, where k denotes the position of the current item, and *i* denotes the lag or the temporal distance to the preceding items – for example, a lag of 2 means that the X_{k-2} item appeared two items ago during study. As we explained before, memory performance for item X_k (denoted by $P(X_k)$) will be worse when the preceding item is weaker (e.g., LF vs. HF). This effect is also not discrete: $P(X_k)$ should be proportional to the strength of item X_{k-1}. These effects should also be cumulative such that $P(X_k)$ should be monotonically worse the more of the preceding items that are weak. This is because more resources will have been spent over time, reducing recovery. The effect of the preceding items X_{k-i} should further increase when the current item X_k is weaker, because the current item is in more need of resources. Finally, the effect of X_{k-1} should be stronger than the effect of X_{k-2} and in general the effect of the preceding items would decrease as the lag increases because more time would have passed, allowing for the recovery and subsequent depletion by other intervening items. These predictions, and the studies in which we found support for them are summarized in Table 7.

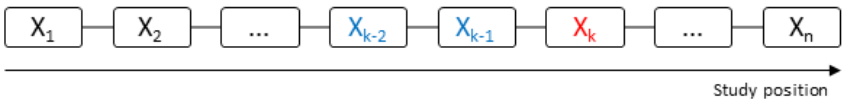

Figure 24. Order of items during a study list

It is important to note that the strength of these effects will depend on the degree to which resources are depleted, which varies as a function of presentation rate, instructions and effort, nature of the material, power issues, etc. It will also depend on the study strategies used by the participants. For example, if for some reason participants decide to study primarily the HF items to maximize their performance, especially if they find the LF items too hard, then the effects are unlikely to appear, or at least be diminished. The discrete effect of prior item strength is the simplest prediction and likely will occur most often, especially when the current item is weak.

We present evidence that supports the predictions shown in Table 7 by reanalyzing nine published studies and by fitting a SAC model to each. The most straightforward variable to analyze is word frequency, but the effects are likely to occur with any variable that affects how many resources are depleted in processing the preceding study items. Of the nine studies, five studies used word frequency as a factor (Healey & Kahana, 2016; Cox et al., 2018; Diana & Reder, 2006; Reder et al., 2002; Ward et al., 2003), two studies manipulated how many times a studied item had already been presented (Buchler et al., 2008; Aue et al., 2017) and two directed-forgetting studies manipulated whether each item should be remembered or forgotten (Marevic et al., 2017; Popov et al., in press).[19]

Table 7. Predictions about the effects of prior item strength

#	Predictions	Studies								
		Diana *	**Ward**	**Buchler**	**Aue**	**Reder**	**Marevic**	**Popov**	**Cox**	**PEERS**
1	**Discrete effect of prior item strength** *Example:* $P(X_k)$ is worse when X_{k-1} is LF	0	+	+	+	0	+	+	NA	NA
2	**Continuous effect of prior item strength** *Example:* $P(X_k)$ is proportional to $freq(X_{k-1})$	NA	NA	+	NA	NA	NA	NA	+	+
3	**Cumulative effect of prior item strength** *Example:* $P(X_k)$ is worse when more of the preceding items are LF	+	0	+	0	+	+	+	NA	NA
4	**Interaction between prior and current item strength** *Example:* The effect of $freq(X_{k-1})$ should be stronger when X_k is LF	+	+	+	+	0	0	0	NA	NA
5	**Interaction between prior item strength and lag** *Example:* The effect of $freq(X_{k-i})$ should decrease as the lag *i* increases	+	0	+	0	0	+	+	+	+

Note. +: effect found in study; 0: null effect; -: effect found in opposite of predicted direction; NA: prediction could not be tested. Diana = Diana & Reder (2006), Ward = Ward et al (2003), Buchler = Buchler et al (2008), Aue = Aue et al (2017), Reder = Reder et al (2002), Marevic = Marevic et al (2017), Popov = Popov et al (2018), Cox = Cox et al (2018), PEERS = Penn Electrophysiology of Encoding and Retrieval Study (Healey & Kahana, 2016)

A. Data analysis

For all results reported below we analyzed the accuracies and RTs via logistic and linear mixed-effects regression models (Baayen, Davidson, & Bates, 2008). We excluded incorrect responses from analyses of RTs. Random effects were determined through restricted likelihood ratio tests

[19] We are grateful to all the researchers who shared their datasets with us. The modeling code, data and analyses scripts are available at https://github.com/venpopov/prior-item-effects

and all final models included varying intercepts for subjects and individual words/word pairs (i.e., subjects and items differ in their overall accuracy and RT estimates). We inferred the significance of each effect based on likelihood ratio tests and AIC comparisons of the regression models that contained the effect in question with identical models that lacked this contrast.

B. Experiments

1. Diana & Reder (2006) – worse memory for items that follow LF items during study

Diana & Reder (2006) performed an item recognition test for pictures that were studied with a HF or a LF word superimposed on it. The authors found that pictures paired with HF words were recognized better. Here we examined the probability of recognizing the picture as a function of whether the *preceding* item during study contained a HF or LF word (see Figure 25, left). There was no main effect of preceding word frequency (Δ AIC = 2, $\chi^2(1) = 0.31$, p = .57), but there was a significant interaction between the frequency of the current word and the frequency of the preceding word during study (Δ AIC = -3, $\chi^2(1) = 5.20$, p = .02). A picture was less likely to be recognized when, during the study list, it was preceded by a trial with a LF word, but only when the current word was also LF. Furthermore, we looked at whether the number of preceding LF or HF words mattered. For each test trial, we calculated how many of the preceding study items were consecutively LF or HF. Picture recognition increased as the number of consecutive preceding LF words decreased from 4 to 1, and it increased as the number of consecutive preceding HF words increased from 1 to 4 (Figure 26; Δ AIC = -2, $\chi^2(1) = 3.62$, p = .038). Finally, we looked at whether the effect of the prior item's frequency would decrease as the lag between it and the current item increased. We limited the analyses only to current LF items, since HF items were not affected by the frequency of the item at the immediately preceding position. As shown on the right panel of Figure 25, the effect was biggest at lag 1 (odds ratio = 1.90, z = 2.216, p = .027), medium at lag 2 (odds ratio = 1.69, z = 1.78, p = .075), and smallest at lag 3 (odds ratio = 1.58, z = 1.57, p = .12). We could not test Prediction 2 with this dataset, because word frequency did not vary continuously. Figure 25 and Figure 26 also show the fit of the SAC model. The model does a good job capturing all effects we described using just a single recovery rate parameter, an episodic retrieval threshold parameter and episodic noise parameter to fit all the data.

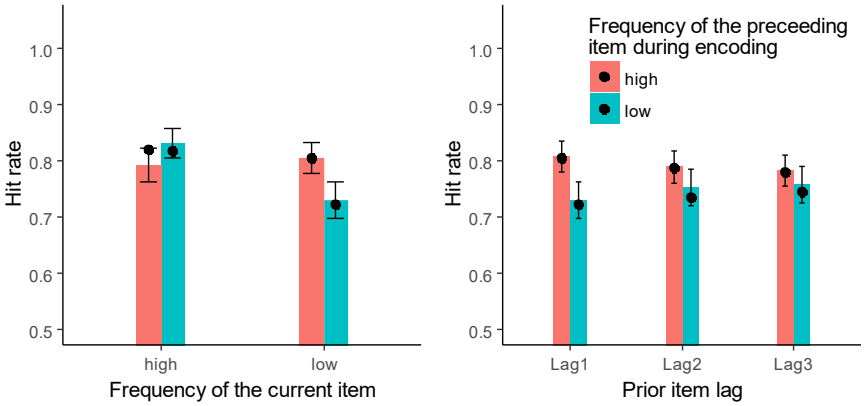

Figure 25. Reanalyzed data from Diana & Reder (2006; bars) and new SAC model fits (dots). Left – hit rate for pictures studied with a superimposed high or low frequency word, depending on whether they were preceded by by a high or low frequency word during study; Right – hit rate for LF words depending on whether they were preceded by a high or low frequency word during study at lag 1, 2 or 3.

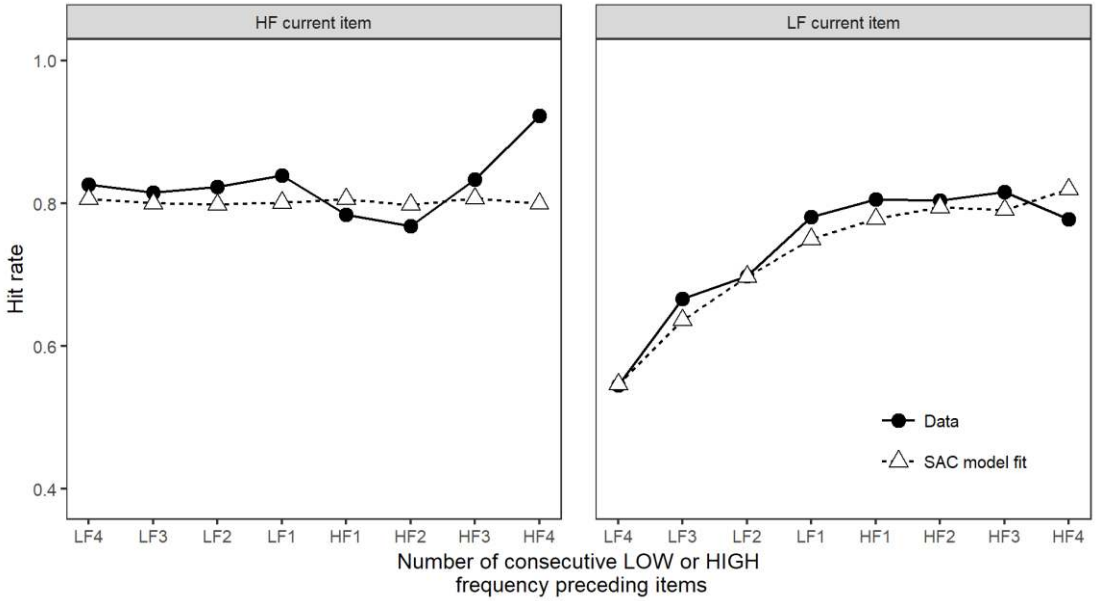

Figure 26. Reanalyzed data from Diana & Reder (2006) and SAC model fits. Hit rate for pictures as a function of how many previous pictures during the study list had low or high freqency words superimposed on them.

2. Ward et al (2003) – rehearsal borrowing cannot explain the preceding frequency effect

Can the prior item effects we found in Diana & Reder (2006) be explained by differences in rehearsal between item with HF and LF words? It might be the case that because LF words are more difficult to remember, participants continue to rehearse them when the subsequent item appears, which limits their ability to process and store the subsequent item. To address this question, we tested whether we would find the same effects when reanalyzing the data from a mixed-list free recall experiment (Ward et al; 2003). Ward and colleagues asked participants to rehearse words out loud, and they recorded the number of times each word was rehearsed. This allows us to test whether rehearsal borrowing can explain the effect of preceding item frequency. The reanalyzed results for Ward's Experiment 3 are shown in Figure 27a. The frequency of the items preceding the current item during the study list had a main effect on recall for the current item – words preceded by LF words during study were less likely to be recalled than words preceded by HF words (Δ AIC = -8, $\chi^2(1) = 9.56$, $p = .002$). As with the Diana & Reder (2006) analysis above, the effect of preceding item frequency was present only when the current item was an LF word (Δ AIC = -3, $\chi^2(1) = 5.07$, $p = .024$). These results could not be accounted for by differential rehearsal – there were no differences in the number of rehearsals depending on the frequency of the prior study item ($p > .3$, Figure 27b). Figure 27 also shows the fit of the SAC

model to the preceding item frequency effects (see Figure 17 for the mixed-vs-pure model fit of the same data).

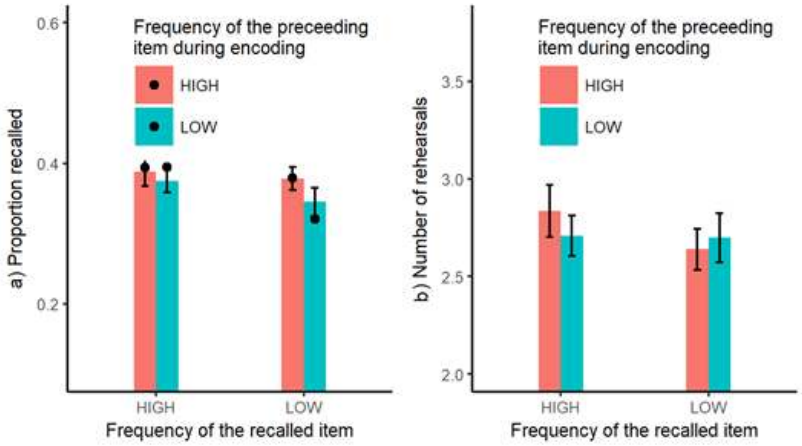

Figure 27. Reanalysis of Ward et al (2003) Dots show the fit of the SAC model. a) proportion recall of high or low frequency words depending on whether they were preceded by by a high or low frequency word during study; b) number of rehearsal of high or low frequency words, depending on whether they were preceded by by a high or low frequency word during study. Error bars represent 95% CI

3. *Cox et al (2018) and PEERS – continuous frequency effect of frequency on memory for the following study item*

Is the effect of preceding item frequency continuous? We explored this question by reanalyzing data from Cox et al. (2018), who asked participants to perform five different memory tasks all using the same stimuli: item recognition, associative recognition, cued recall, free recall and lexical decision. We focus on the first four tasks, because the lexical decision task did not involve study lists. In all lists, participants studied word pairs without knowing the nature of the subsequent test. After studying a given list, they were given one of the four tests described above. Word frequency was not a factor in this experiment, but the frequency of the words still varied, allowing us to test whether memory for the current item X_k will be affected continuously by the frequency of the prior word. We calculated the frequency for each trial by averaging the log SUBTLEX frequency of each word within a pair (van Heuven et al, 2014).

Figure 28 shows the hit rate (in recognition tasks)/recall accuracy (in recall tasks) and the RTs for each memory task as a function of the preceding study item's frequency (binned in 20 frequency groups with equal number of observations). Performance in all tasks improved continuously as a function of the preceding study item frequency, $\Delta AIC = -5$, $\chi2\ (1) = 6.63$, $p = .01$. There was no significant interaction between prior item frequency and task type, $\Delta AIC = 5$, $\chi2\ (3) = 0.99$, $p = .80$. Finally, we found that the effect of prior item frequency decreased with the lag between the current item and the prior item during study, which was also predicted by our model (Figure 29a, b, respectively) – the effect was biggest at lag 1 (odds ratio for 1 log unit increase in frequency = 1.08, $z = 2.58$, $p = .009$), medium at lag 2 (odds ratio for 1 log unit increase

in frequency = 1.03, z = 0.86, p = .388), and smallest at lag 3 (odds ratio for 1 log unit increase in frequency = 0.99, z = -0.51, p = .61). In summary, the effect of prior item frequency is continuous and its impact diminishes as the study lag between the current and the prior item increases.

We fit the SAC model to Cox et al.'s accuracy data (we did not fit RTs, but the behavioral data for RTs is shown to illustrate the absence of a speed-accuracy trade-off). Since participants did not know the nature of the next test when studying the word pairs, we used a single model to fit performance in all tasks. This involved only a single parameter for recovery rate, retrieval threshold and retrieval noise for all tasks. As shown in Figure 28, the model did a good job fitting the overall hit rates across the four memory tasks. Given that the nature of the test was not known in advance, the model treated all study sessions equivalently and any differences in performance must be due to how much activation reaches the episode nodes during retrieval. Specifically, for free recall, episode nodes only receive activation from the context node; for cued recall, activation spreads from both the context node and the activated cue (word) node; for associative recognition activation is sent to the episode node from the context node and the two cue nodes (See Figure 4 for an illustration). For these three tasks the amount of activation of the episode node increases with the number of activation/cue sources during retrieval. The single recognition task is slightly different – while the episode node receives just as much activation during single recognition as it does during cued recall, responses in the single recognition task can be based on either the activation of the episode node (resulting in Remember responses in a Remember-know paradigm) or based on activation of the semantic node of the cue (resulting in Know responses; see Reder et al., 2000 for greater detail). Since Cox et al. (2018) used an old-new instead of a remember-know paradigm, we assume both nodes' activations contribute to the Hit rate (Diana et al., 2006).

Figure 29 and Figure 30 show the fits to the preceding item frequency effects. Figure 29 shows that the model captures well the slope of the frequency effects. The intercept for each task is not fit perfectly, which is due to the fact that we used a single retrieval threshold for all tasks and instead depended on inherent differences in activation levels between retrieval tasks to capture the overall hit rates. Figure 30 shows that the model also captures the fact that the effects due to preceding items decrease with lag between the target and the relevant preceding item.

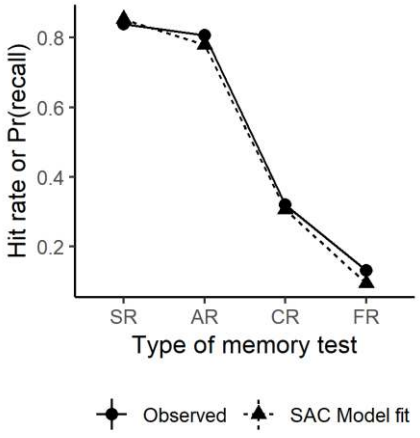

Figure 28. Overall results in Cox et al (2018) and the SAC model fit to the four different test types. SR – single item recognition; AR – associative recognition; CR – cued recall; FR – free recall.

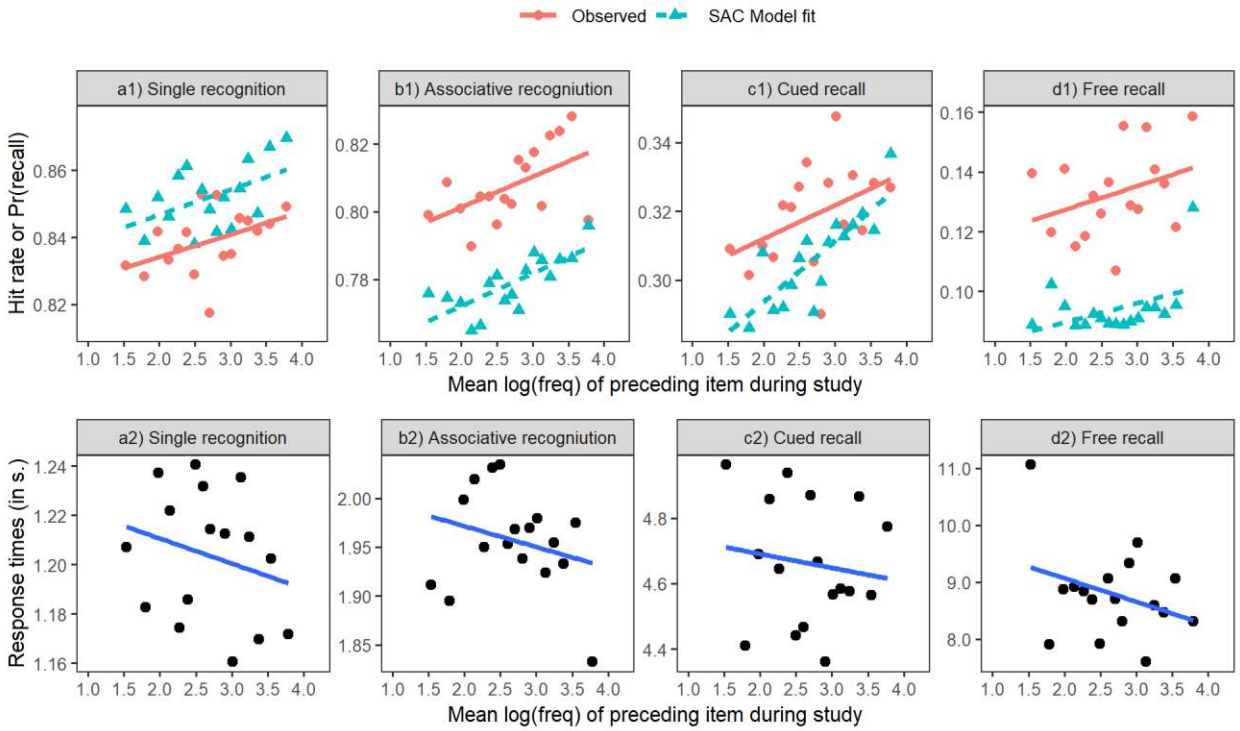

Figure 29. Reanalysis of Cox et al (2018) and SAC model fits to the accuracy data – Hit rates or probability of recall (top panels) and RTs for correct responses (bottom panels) as a function of the mean word frequency of the word pair that preceded the current pair during study and task type - a) Single recognition, b) Associative recognition, c) Cued recall, d) Free recall. Frequencies were binned into 20 bins of equal size and points represent the mean in each bin. Variability in model predictions reflects the particular mix of trial sequences, since model predictions are derived by simulating activation values given the actual trial sequences. Lines show the best fitting regression line to the data and the model.

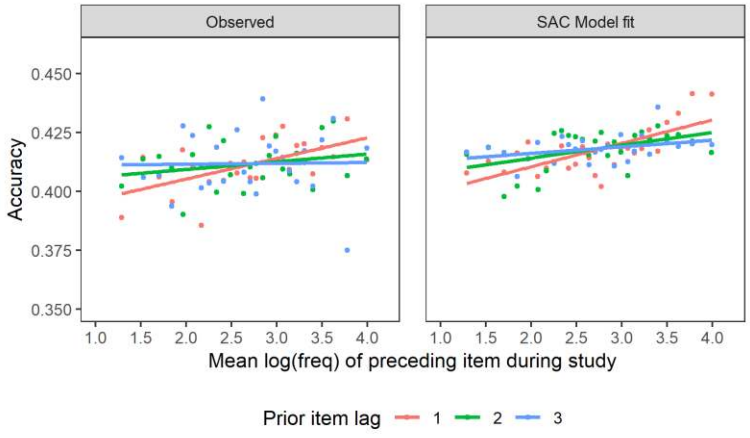

Figure 30. Reanalysis of Cox et al (2018) and SAC model fit – Accuracy (hit rate for recognition tasks or probability of recall for recall tasks) over all tasks depending on the frequency of the items that preceded the current item during study and their lag (e.g., how many trials earlier they occurred). Left – empirical data; Right – simulated data from the SAC model. Variability in model predictions reflects the particular mix of trial sequences, since model predictions are derived by simulating activation values given the actual trial sequences. Lines show the best fitting linear regression line to the data and the model.

To replicate the continuous effect that we found in Cox et al. (2018), we also reanalyzed data from the Penn Electrophysiology of Encoding and Retrieval Study (PEERS), a large-scale multi-session experiment on free recall in which participants studied words presented one at a time (Healey & Kahana, 2016; Lohnas & Kahana, 2013). Here we analyze data from the 230 participants who completed Experiment 1[20]. In each of seven sessions, participants studied 16 lists of 16 words presented one at a time. After each list, participants had to immediately perform a free recall test for the preceding list (see Healey & Kahana, 2016 or Lohnas & Kahana, 2013 for full procedural details).

The effects of preceding item frequency and its interaction with the lag to the preceding item in PEERS were remarkably similar to those we found in Cox et al. (2018). Figure 31 shows free recall accuracy and RTs as a function of the preceding study item's frequency (binned in 20 frequency groups with equal number of observations). Recall accuracy improved when the preceding item during study was of higher frequency, $\Delta AIC = -41$, $\chi 2\ (1) = 42.95$, $p < .001$. Similarly, RTs were faster when the preceding item was of higher frequency, $\Delta AIC = -10$, $\chi 2\ (1) = 12.47$, $p < .001$. We also found that the effect of prior item frequency decreased with the lag between the current item and the prior item during study (Figure 32) – the effect was biggest at lag 1 (odds ratio for 1 log unit increase in frequency = 1.029, $z = 4.856$, $p < .001$), medium at lag 2 (odds ratio for 1 log unit increase in frequency = 1.013, $z = 2.149$, $p = .032$), smallest at lag 3 (odds ratio for 1 log unit increase in frequency = 1.010, $z = 1.728$, $p = .08$) and absent at lag 4 (odds ratio for 1 log unit increase in frequency = 1.002, $z = 0.481$, $p = .63$). Thus, our analyses of the PEERS dataset replicated the effects we found in Cox et al (2018).

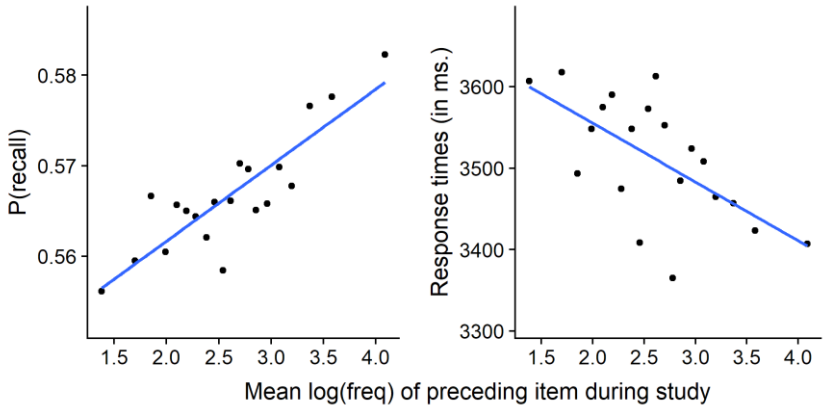

Figure 31. Reanalysis of PEERs and SAC model fits to the accuracy data –probability of recall (left) and RTs for correct responses (right) as a function of the mean word frequency of the word pair that preceded the current pair during study. Frequencies were binned into 20 bins of equal size and points represent the mean in each bin. Lines show the best fitting linear regression line to the data and the model.

[20] The results were equivalent when including Experiment 2 and 3, but since not all participants completed all experiments, we focus on Experiment 1 for simplicity.

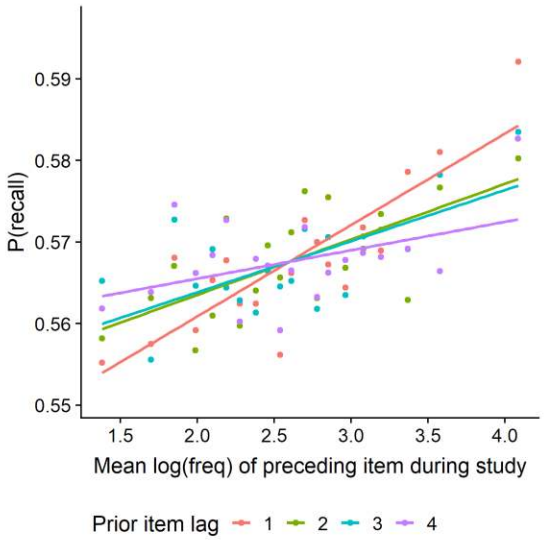

Figure 32. Reanalysis of PEERs and SAC model fit – Probability of recall depending on the frequency of the items that preceded the current item during study and their lag (e.g., how many trials earlier they occurred). Left – empirical data; Right – simulated data from the SAC model. Variability in model predictions reflects the particular mix of trial sequences, since model predictions are derived by simulating activation values given the actual trial sequences. Lines show the best fitting linear regression line to the data and the model.

4. *Buchler et al (2008) – number of study repetitions of an item affects subsequent items in the same way as their normative frequency*

The next two studies that we re-analyzed involved manipulations of item strength by varying the number of repetitions during study. According to SAC, when some items are repeated multiple times on a study list their strength will be higher with each repetition and each subsequent encoding should require fewer resources. That means that there should be more resources available to store other items that follow items were repeated multiple times.

Buchler et al. (2008) performed an associative recognition experiment, in which, among other conditions, they manipulated how many times a specific word pair was repeated (either 1 or 5). For each study trial, we calculated whether it was preceded by a word pair that had been presented 1, 2, 3, 4 or 5 times already. For simplicity, we collapsed the study trials into two groups – we examined whether a word pair X_{k-1} appeared for the first time (weak pair) or whether it had been repeated multiple times (strong pair). Figure 33 shows the results of this analysis. Hits for studied pairs were higher and false alarms to recombined pairs were lower when the target had been studied right after a repeated pair, Δ AIC = -2, $\chi2(1) = 3.92$, $p = .047$. Importantly, the effect of the prior item strength interacted with the current pair's strength: the advantage of following a strong pair was bigger when the current item was weak (studied only once), Δ AIC = -14, $\chi2(1) = 16.35$, $p < .001$. The effect of a prior item's strength decreased as the lag between study positions increased – the effect was biggest at lag 1 (odds ratio = 1.68, $z = 4.134$, $p < .001$), medium at lag 2 (odds ratio = 1.11, $z = 2.914$, $p = .004$), and smallest at lag 3 (odds ratio = 0.94, $z = 1.465$, $p = .143$). Performance was also a function of how many of the preceding items during study were weak or

strong (Figure 34, right). Finally, while for the previous analyses we dichotomized pairs into either weak (one repetition) or strong (multiple repetitions), we also found that the effect of the preceding item's strength was continuous, such that hits were higher when the preceding trial during study contained a pair that was repeated more (Figure 34, left). In summary, we found support for all five predictions with manipulated frequency just as we had with normative (natural) word frequency. The figures also show the fit of the SAC model, which did a good job accounting for all of the effects.

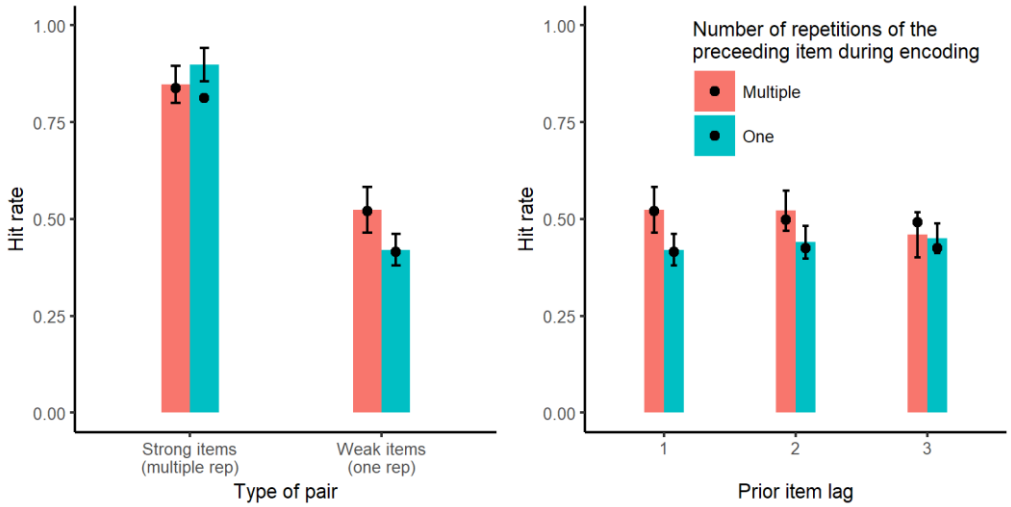

Figure 33. Reanalysis of Buchler et al (2008) and SAC model fits (dots). Left panel – Hits for repeated and non-repeated items depending on whether they were preceded during study by a repeated or a non-repeated pair. Right panel - hit rate for weak items depending on whether the items were preceded by a repeated or a non-repeated pair during study at lag 1, 2 or 3. Error bars represent 95% CI

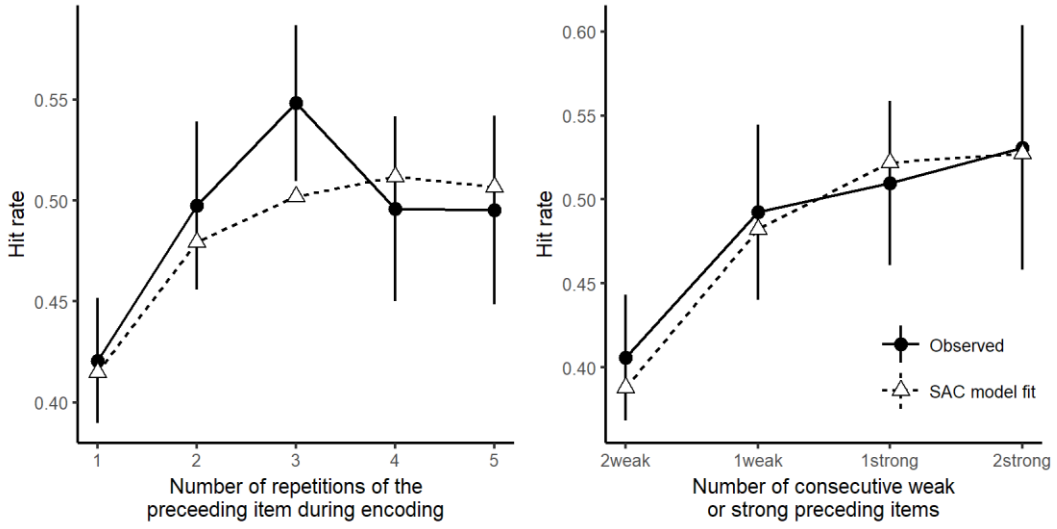

Figure 34. Reanalysis of Buchler et al (2008) and SAC model fits. Left panel – hits for the current item depending on how many times the preceding item was repeated during study. Right panel – hits as a function of how many trials during the study list had weak (one repetition) or strong (multiple repetition) pairs preceding the test pair. Error bars represent 95% CI

5. *Aue et al (2017) – is repeating one word of a pair sufficient to facilitate memory for the following study item?*

Does the whole pair need to be repeated, in order for the following item to be better remembered? Our model would predict that even repeating only one word of the pair across pairs would cause less resource depletion, leaving more resources for the subsequent item (though the effect would be weaker). We explored this question by reanalyzing data from three experiments concerned with proactive facilitation (Aue, Criss, & Novak, 2017). In each of their cued-recall experiments, participants studied two lists of word pairs and some of the pairs on the second list shared a word with pairs in the first list. The authors found that cued recall for List 2 was better for those pairs that share words with List 1 compared to pairs with cues unique to list 2. The authors argued, similar to our position, that when a cue was repeated from List 1 to List 2, it was more familiar, making it easier to associate to a new target.

As in Buchler et al. (2008), we looked at performance for the current pair depending on whether it was preceded during study by a pair that was presented for the first time (new), or whether it was preceded by a pair that contained a previously seen cue word from List 1. We combined the data from Experiments 1, 2 and 4 (Exp. 3 tested List 1 performance and was thus irrelevant). Figure 35 shows the results and the corresponding SAC model fit. Cued recall was higher when the pair was preceded during study by a pair with a repeated cue (Δ AIC = -3.1, $\chi2(1) = 5.10$, $p = .024$ for the main effect of preceding cue type), but only when the current item was new (Δ AIC = -3, $\chi2(1) = 4.74$, $p = .029$ for the interaction between preceding and current cue type). Thus, we replicated the main effect we found in Buchler et al. (2008). It is likely that the effects are weaker, because in this study only one of the words was repeated, not the entire pair, and it was repeated only one time, not up to five times as in Buchler et al. (2008).

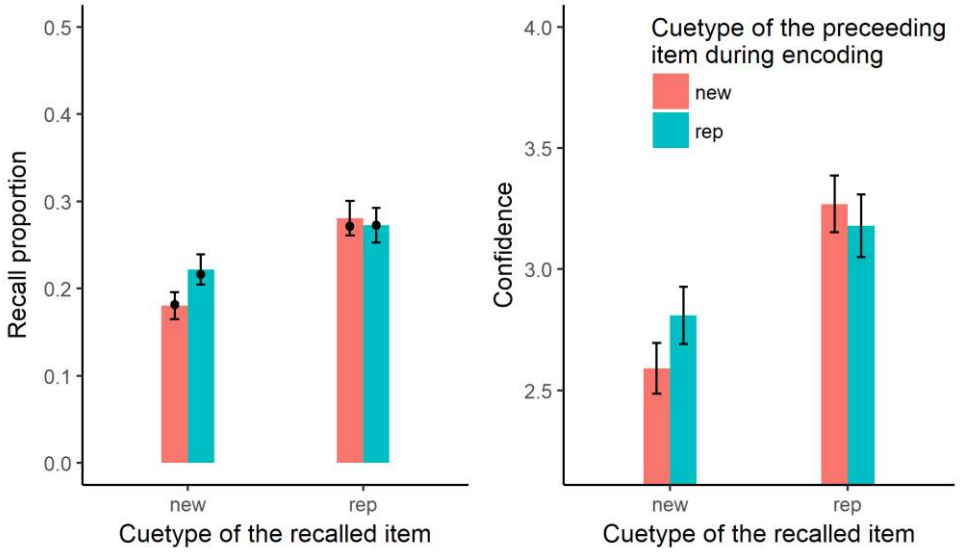

Figure 35. Reanalysis of Aue et al (2017) and SAC model fit (dots). Cued-recall probability (left) and confidence ratings (right) for pairs with a new or a repeated cue, depending on whether they were preceded during study by a pair with a new or a repeated cue. Error bars represent 95% CI

6. Reder et al (2002) – extending the predictions to experimentally manipulated frequency

Throughout this text, we discussed the benefits of drawing inferences from training studies that experimentally manipulate stimulus familiarity prior to the task of interest, as compared with studies that use normative word frequency (a quasi-experimental variable). Would the effects of the preceding item familiarity show up in such training studies as well? We reanalyzed data from Reder et al. (2002) who trained participants to learn pseudowords over a period of five weeks with some pseudowords presented six times more often than other pseudowords. The primary task was free recall of the just presented pseudowords on a list. This training occurred three times per week. Once at the start of each week, before the standard free recall training, participants were given a list for which the task was episodic pseudoword recognition that required the participant to discriminate the items just studied on the list from other pseudowords being trained in the study but not presented.

We compared recognition memory accuracy for items that were preceded by HF or LF pseudowords. There was evidence of a speed-accuracy trade-off (Reed, 1973), such that the responses to the conditions we expected to be worse were less accurate but also faster. To finesse the resulting interpretation problem, we report analyses done on inverse efficiency scores (recognition test RTs divided by average recognition accuracy, separately for each condition; Townsend & Ashby, 1983). The main effect of the immediately preceding item was not significant (Δ AIC = 0, $\chi2(1) = 2.00$, $p = .15$), and there was no interaction with the current item frequency (Δ AIC = 2, $\chi2(1) = 0.36$, $p = .54$). However, as shown in Figure 36, there was a cumulative effect of the number of prior LF or HF items: Pseudoword recognition improved as the number of consecutive preceding LF decreased from 3 to 1, and it increased as the number of consecutive preceding HF words increased from 1 to 3, (Δ AIC = -9, $\chi^2(1) = 10.85$, $p < .001$). Thus, even though the immediately preceding item effect was not significant, we found support for prediction 3 just as in Diana & Reder (2006).

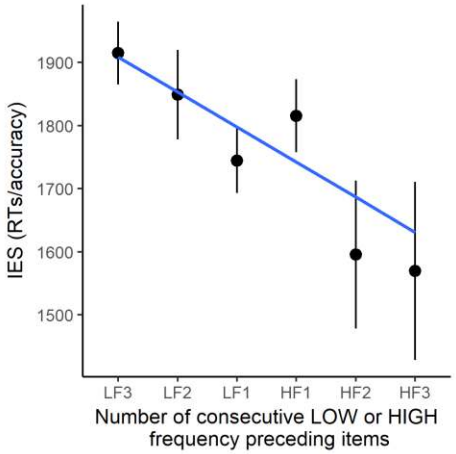

Figure 36. Reanalysis of Reder et al (2002). Memory performance in inverse efficiency (IES = RTs/accuracy) as a function of how many trials during the study list had low or high freqency words preceding the test word. Error bars represent 95% CI

7. *Marevic et al (2017) – better memory after instructions to forget the preceding study item.*

While our focus has been primarily on differences in strength due to word-frequency and experimental familiarization, SAC predicts that similar preceding item effects should occur with any variable that affects how much resources are depleted. To test this idea, we re-analyzed data from a Directed Forgetting (DF) study (Marevic, Arnold, & Rummel, 2017). In the item-based version of the DF paradigm, participants view a sequence of items and they are tested only on items followed by a "remember" cue during study (TBR items) and participants are supposed to forget the items followed by a "forget" cue (TBF items). SAC's prediction is that memory for items following TBR items should be worse because more WM resources will be consumed trying to learn the previous item compared to when the previous item was TBF.

The full analysis and results are described in Popov et al. (in press), but we will briefly summarize those results. The analysis of preceding study items showed that cued-recall and free-recall paired-associate memory for items presented on study trial *k* was better when study trial *k-1* was TBF, rather than TBR. There was also a cumulative effect – when more of the previous trials (k-3, k-2, k-1) were to-be-remembered, memory for the item on trial k was worse. Finally, the effect of preceding item cue type decreased as the study lag between the current item and the prior item increased (e.g., the item on trial *k-3* had a weaker effect than that on trial *k-2*, which in turn had a weaker effect than the item on trial *k-1*).

8. *Popov et al (2018) – could the directed forgetting results be explained by rehearsal or attentional borrowing?*

It is possible that when people study an item, they continuously rehearse or reactivate the memory traces for the preceding items (Camos, Lagner, & Barrouillet, 2009), and that this rehearsal or attentional borrowing is greater when more previous items had to be remembered. Popov et al. (in press) did not find support for that explanation in a dual task study designed to test that explanation. Participants tried to learn paired-associates in one of four conditions - a control single-task condition, equivalent to Marevic et al. (2017), a rehearsal suppression dual task condition, a divided attention dual task condition and a combined rehearsal suppression plus divided attention dual task condition. All of the effects described above were replicated regardless of whether attention was divided or rehearsal was suppressed. In other words, the effects of the prior item type (TBR vs TBF) was as strong when there was no ability to rehearse.

C. Meta-analysis

While the majority of predictions were supported by the majority of the studies (23 out of 31, Table 7), it is worth asking how reliable the discovered effects are overall. Since we have access to the trial level data for each study, we performed an Individual Participant Data (IPD) meta-analysis, which has a number of advantages over traditional aggregate data meta-analyses (Tierney et al., 2015). In an IPD meta-analysis, rather than analyzing the aggregate effect sizes for each study, one can perform the same mixed-effects logistic regression analyses that we reported for

each study on the combined data from all studies. In addition to adding random intercepts and slopes for each participant, the analysis includes random intercepts and slopes for the effects of interest for each study. All five predictions were confirmed in the meta-analysis. There was an overall main effect of prior item strength, such that memory across all studies was worse when the preceding study item was weak rather than strong, Δ AIC = -9, $\chi2(1) = 11.21$, $p < .001$. The effect of the prior item strength interacted with the current item strength, such that it was larger when the current item was weaker, Δ AIC = -2, $\chi2(1) = 3.88$, $p = .048$. The effect of the prior item strength decreased as the study lag between the current and the prior item increased, Δ AIC = -13, $\chi2(1) = 14.52$, $p < .001$. The effect of the prior item strength was cumulative, such that a greater number of weak items in a row lead to even worse memory for the current item, Δ AIC = -13, $\chi2(1) = 14.99$, $p < .001$. Finally, the effect of prior item strength was continuous Δ AIC = -36, $\chi2(1) = 38.01$, $p < .001$.

D. Discussion

Evidence from nine studies support the novel prediction that memory for a given item is affected by the strength of the preceding items during the study phase. Specifically, memory for item X_k studied on trial *k* was better when the preceding item during study, X_{k-1}, was stronger and thus required fewer resources to be processed. This occurred when the preceding item was a HF rather than a LF word or word pair (Diana & Reder, 2006; Ward et al., 2003; Cox et al., 2018; PEERS – Healey & Kahana, 2016), was a pseudo-word that was experienced more frequently in training prior to the study list (Reder et al., 2002) or was a word or word pair that was repeated on multiple trials (Buchler et al., 2008) or lists (Aue et al., 2017). A similar memory advantage occurred when the preceding word pair was supposed to be forgotten, rather than remembered, suggesting that to-be-remembered word pairs deplete more resources relative to to-be-forgotten word pairs (Marevic et al., 2017; Popov et al., in press). The carry-over effect was a parametric function of word frequency (Cox et al., 2018; PEERS - Healey & Kahana, 2016), it accumulated when more of the previous items during study were weaker (Diana & Reder, 2006; Buchler et al., 2008; Reder et al., 2002; Marevic et al., 2017; Popov et al., in press), it decreased when the lag between the current and the prior item increased (Cox et al., 2018; Diana & Reder, 2006; Buchler et al., 2008; Marevic et al., 2017; PEERS - Healey & Kahana, 2016; Popov et al., in press), and it was bigger when the current item was weaker itself (Buchler et al., 2008; Diana & Reder, 2006; Ward et al., 2003). Importantly, these results cannot be due to other mechanisms such as rehearsal, distinctiveness, or attentional borrowing (Popov et al., in press). These results provide strong support for the claim that memory formation depletes a limited pool of resources as a function of the current strength of items, and that these resources recover over time.

V. General Discussion

The resource-depletion theory can account for many effects of pre-existing item strength on learning and WM (summarized in Tables 3, 5, 6 and 7). We demonstrated that it is easier to form

new memories for items with stronger current nodes in LTM, such as high frequency (HF) words or frequently exposed novel stimuli. Such items are also easier to maintain in WM, and we discounted alternative explanations for these results. Furthermore, we showed that the strength of one item interacts with surrounding items; we also provided evidence from pure-vs-mixed list comparisons and reanalyzed 9 existing datasets that showed memory for one item depends on the strength of the items that immediately preceded it during study. The current theory accounts for these results by positing a limited resource that is used in processing, storage and maintenance, a resource that recovers slowly over time. According to the theory, the encoding advantage for stronger items occurs because they require fewer resources for their processing and, as a result, more resources are left for binding them to the study context and to one another. The model simulations provided good fits to many of the patterns discussed in the paper. In the remainder of this discussion we will consider some difficulties in providing evidence for the current proposal, describe some additional phenomena that the theory might explain, and we will conclude by discussing the concept of resources and its usefulness in explaining memory performance.

A. Partial matching's role in WM resource depletion

Replication has recently become an important issue in the field of psychology (Pashler & Wagenmakers, 2012). While the WM familiarity advantage has been demonstrated with quasi-experimental studies (reviewed here and also in Reder et al., 2007; also see Xie & Zhang, 2016, 2017b), based on our experience, it is quite difficult to demonstrate *experimentally* that familiar items deplete fewer WM resources, and we believe it is important to discuss the necessary conditions under which training studies like Reder et al. (2016) and Shen et al. (2018) can successfully demonstrate the effect. Several initial unpublished training experiments from our lab failed to find the effects demonstrated in Reder et al. (2016). That difficulty is important to understand because (a) without explaining the necessary conditions for getting the effect, there would be failures to replicate, leading to invalid rejections of the theory; and (b) the explanation for the difficulty is closely tied to the theory.

These previous experiments, which failed to find a familiarity advantage, used a training procedure that was similar to the one used by Reder et al. (2016) – participants were familiarized with previously unfamiliar Chinese characters at either low or high frequency in a visual search task, and then they had to perform a cued recall task in which two characters were associated with an English word. There are two key differences between these earlier studies and Reder et al. (2016) and Shen et al. (2018) that we think are crucial for understanding why the experiments failed to find support for the theory. First, in the visual search task, distractors in the failed studies were selected at random, rather than from a set of highly similar characters. Second, in the cued recall task, each character was presented in only one triplet (A-B-Chair; C-D-Sky), rather than in two separate triplets (A-B-Chair; A-C-Sky, etc).

Why do we think these two differences were responsible for not finding a familiarity advantage? The main issue concerns *Partial Matching*, a natural adaptation in situations where WM resources are scarce. Partial matching operates regularly in cognitive processing, especially

when a person's WM resources are exhausted. People can only attend to so much and they focus selectively, tending not to notice the rest. This phenomenon underlies the Moses Illusion (e.g., Kamas, Reder, & Ayers, 1996; Reder & Kusbit, 1991), change blindness (Simons & Levin, 1997), spurious Feeling of Knowing in arithmetic tasks (Reder & Ritter, 1992; Schunn et al. 1997) and *trivia game-show* experiments (Reder, 1987). These phenomena share the property that there tends to be too much information for a person to focus on simultaneously, leading them to accept as a match information that only partially overlaps with the contents of memory.

So *how or why* does partial matching make it difficult to demonstrate our theoretical position? In order to preserve WM resources, partial matching could occur at the character level or at the level of features within a character. In the original unpublished experiments, because each character was presented in only one triplet, participants could perform the cued recall task even if they only encoded one of the two characters in a pair. Given that LF characters were more challenging, more partial matching *at the character level* would occur for those trials and would dilute the effect. To fix this issue, subsequent experiments involved using each character as part of two pairs (each with different English words) on a given study list (e.g. the character A was present in both the A-B-CHAIR and A-C-SKY triplets). This change forced participants to bind both characters to the target word, preventing partial matching, and the main result emerged – performance was better for HF pairs compared to LF pairs. However, the results still showed some odd patterns in a zero frequency exposure conditions – counterintuitively, triplets with completely novel characters (i.e., a zero frequency exposure condition) were learned better than the LF pairs[21].

Why were the zero frequency pairs outperforming low-frequency pairs? We realized that participants could also partial match *at the feature level* within a character as well, finding features in characters that were already strong chunks. When people only encode those parts of a novel stimulus that involve previously known chunks, rather than (in this case) the entire Chinese character, there are fewer demands on the WM resource. However, this short-cut means that a complete representation of the character is not created and defeats the goal of controlling the familiarity of the stimulus. With characters that were studied a few times per week, the beginnings of new chunks representing the entire character would be formed but would not be particularly strong. Those lists consisting of zero frequency characters would not have the WM demands of the LF characters, because participants likely partial-matched to strong, familiar chunks. Figure 37 illustrates how someone might encode personally strong chunks such as the symbol π, the number *4* or the letter *k* in those unfamiliar stimuli instead of the entire characters.

[21] Nelson & Shiffrin (2013) reported a similar effect. They found a non-monotonic trend for single character learning, such that zero frequency did better than low frequency (they did not train pairs nor associate them to English words; their study involved a recognition test).

Figure 37. Examples of three Chinese characters and potential chunks (highlighted in grey) that might be encoded (π, 4 and k) and used for partial matching.

The question was how to get subjects to not rely on feature-level partial matching when the chunks were not strong. Feigenbaum and Simon (1984) argued that people only encode those features required for discriminating among categories, and that explains why, for example, people who used pennies every day of their lives for decades failed to discriminate the correct representation of a penny from plausible foils (Nickerson & Adams, 1979). With this insight, we made the characters in the training sets visually similar.[22] When fine-grained discrimination between characters was forced in this way, the predicted effects finally emerged.[23] In summary, we suggest that *experimental manipulations of frequency require discrimination training with highly similar foils for the benefits of exposure frequency to emerge.*

B. Accounting for additional phenomena

1. Primacy effects

A natural extension of the resource depletion mechanism discussed so far is to account for primacy effects in list learning, the phenomenon of better memory for items presented early in a list. Traditional accounts of the primacy effect posit that early list items are rehearsed more often (Rundus, 1971). We do not find this explanation satisfactory because of two findings. First, the number of rehearsals does not necessarily predict long-term retention (Craik & Watkins, 1973; Glenberg, Smith, & Green, 1977). Second, even when rehearsal is suppressed, a primacy effect remains, albeit smaller in size (Marshall & Werder, 1972; Howard & Kahana, 1999). These findings motivate a new explanation consistent with the current theory – that the processing of each item depletes WM resources and that the primacy effect is due to fewer resources being available on each subsequent trial after the first.

---

[22] These were selected by Xiaonan Liu, a native Chinese speaker and Reder's student at the time

[23] We know that the effect emerged due to the higher visual similarity of the distractors, because the effects were absent in an otherwise identical version of the experiment that contained a bug in the code causing distractors to be randomly assigned instead

A number of findings are consistent with this explanation. High WM capacity individuals show a reduced primacy effect (Unsworth & Spillers, 2010), which could be due either to having more resources or to having a faster resource recovery rate. Furthermore, since WM resources recover as a function of time, we expect that the primacy effect should increase as the presentation rate gets faster, because WM resources would have recovered less between items. Consistent with this prediction, a stronger primacy effect is found with immediate serial recall with a rapid serial visual presentation paradigm with 9 words per second compared to 1 word per second (Coltheart et al., 2004; Neath & Crowder, 1996; Foster, 1970). The effect is robust – it occurs with forward serial recall (Coltheart et al., 2004), backward serial recall (Neath & Crowder, 1996), serial order recognition test (Coltheart et al., 2004), yes-no probe recognition test (Potter et al., 2002).

Our explanation of the primacy effect is similar to attentional gradient models, according to which during encoding less and less attention is paid to each additional item (for a review, see Oberauer, 2003). For instance, in the Temporal Context Model (TCM; Howard & Kahana, 2002; Sederberg, Howard, & Kahana, 2008), the primacy effects result from an encoding boost parameter that decays exponentially with each trial. This is a convenient way to fit the data, but is not in and of itself a mechanistic explanation of primacy effects (Sederberg et al., 2008). We propose that the primacy gradient arises due to the same mechanism we have used to account so far for the various frequency and list composition effects – the encoding of items depletes a limited pool of resources, which recover over time.

2. *Better WM memory in experts in their domain of expertise, even for random configurations of stimuli*

One classic result in the memory literature is that experts are much better at memorizing material within their domain of expertise, compared to non-experts (Chase & Simon, 1973; Chiesi, Spilich, & Voss, 1979). While initial studies showed that chess masters and novices do not differ in memorizing ability for *random placement* of chess pieces on a board (Chase & Simon, 1973), later analyses revealed that there is a robust, albeit smaller, advantage of expertise even for random chess positions (Gobet & Simon, 1996). A recent meta-analysis showed that this advantage for random configurations occurs in many other domains such as music, sports or computer programming (Sala & Gobet, 2017). It is possible that experts recognize smaller subsets of chunks even within randomly structured materials, and that these chunks allow them to memorize more of the information (see Gobet, 1998, for a theoretical review; Sala & Gobet, 2017). An additional contributor might be the experts' greater familiarity with individual items from their domain of expertise, e.g., chess pieces and board positions, musical notes, programming commands, or climbing moves. An expert might exhaust fewer WM resources in binding these more familiar items into a random configuration compared to a novice with the same overall WM capacity.

3. Better LTM for people with more WM capacity

Given the exposition so far, perhaps it is not surprising that people with higher WM capacity also perform better on episodic LTM tests (Anguera et al., 2012; Anguera, Reuter-Lorenz, Willingham, & Seidler, 2010; Marevic, Arnold, & Rummel, 2017; Unsworth, Brewer, & Spillers, 2009; Unsworth & Spillers, 2010). Multinomial modeling has revealed that WM capacity correlates with encoding success, rather than with retrieval probability (Marevic et al., 2017). This correlation is one of the key puzzles that a theory of WM must explain, and Oberauer et al. (2016) noted that no current resource-based account of WM can do so. The theory presented in this paper provides a natural account of for this observation – the creation and binding of novel episode nodes depletes a limited resource, and the strength or the probability of forming these nodes depends on the quantity of WM resources available. People with less WM capacity, either due to less total capacity, or due to slower recovery rate, would form fewer or weaker chunks, especially when the demands are greater.

4. Binding problems in old age

The current theory also provides an account of the correlation between WM capacity and learning performance evident in the effects of aging on both WM and episodic learning (Buchler & Reder, 2007; also see Buchler et al., 2011). WM capacity decreases at a steady rate from about 20 years of age onwards (Brockmole & Logie, 2013; Salthouse & Babcock, 1991), and this decline is greater for bindings than for item information (Cowan, Naveh-Benjamin, Kilb, & Saults, 2006; Peterson & Naveh-Benjamin, 2016, 2017). A large array of varied findings from item and associative recognition also indicates that the underlying cause of most age-related memory impairments is a decreased ability to form novel bindings (Ahmad, Fernandes, & Hockley, 2015; Buchler et al., 2011; Chalfonte & Johnson, 1996; Light, Patterson, Chung, & Healy, 2004; although see Bartsch et al., 2019). Our model suggests a causal link between the correlated decline in WM capacity and binding ability with age – forming new bindings requires WM resources and the decrease in WM capacity is responsible for the reduced ability to establish and store such bindings (Buchler et al., 2011).

What makes us think that WM decline is related to the binding deficit? First, neuroimaging evidence is consistent with this claim. There is reduced activity in prefrontal areas involved in WM in old adults when binding is required, e.g., during the encoding of unfamiliar faces (Grady et al., 1995) and word pairs (Anderson et al., 2000; Cabeza et al., 1997), but not during encoding single items (Grady et al., 1995). Second, divided attention during encoding, which reduces the amount of resource available for memory operations, reduces or sometimes eliminates the age difference (Anderson et al., 2000; Craik & Byrd, 1982; Jennings & Jacoby, 1993), and it also leads to a reduction in prefrontal WM-related activity during encoding (Anderson et al., 2000; Shallice et al., 1994). Finally, recent studies are finding that, similar to LTM episodic memory, older adults' WM impairments are greater when they have to maintain item-item or item-context bindings (Peterson

& Naveh-Benjamin, 2016, 2017). This overall pattern of results is consistent with the proposed connection between WM decline and reduced binding ability in old age.

While this explanation can account for many findings in the literature, it is challenged by a recent study by Bartsch, Loaiza and Oberauer (2019). The current theory would predict that increasing WM demands during encoding should impair not only immediate WM performance, but also subsequent episodic memory for the same stimuli. Bartsch and colleagues asked participants to study lists of 2 to 6 word pairs, and then tested their cued-recall performance both immediately after each list, and after a block of five lists. Increasing the set size drastically decreased immediate WM performance; however, subsequent LTM for the same stimuli did not differ depending on the study list set size. While we cannot explain this result, it is surprising given that typically list length does affect cued recall (Murnane & Shiffrin, 1991) and we suspend judgement until this finding is thoroughly replicated.

5. Children and WM

An implication of the argument that WM depends on the strength as well as the number of chunks that must be processed allows us to question the conventional wisdom that children have less WM than adults (Brockmole & Logie, 2013; Gathercole, Pickering, Ambridge, & Wearing, 2004; Nelson Cowan, Ricker, Clark, Hinrichs, & Glass, 2015). The assertion that the size and the strength of the chunks increase from childhood to adulthood invites the speculation that WM resources *only appear to increase* from childhood to adulthood.

There is evidence consistent with this proposal. For example, Chi (1978) demonstrated that children who are expert chess players have a better memory for meaningful chess positions compared to adult novices, despite a worse WM span measured with conventional stimuli. Extending these results, Schneider et al. (1993) found that the child chess experts' memory advantage in immediate recall holds even when compared with adult chess experts. Furthermore, the memory advantage was even greater for random configurations of chess stimuli. Random configurations do not match existing chunks in memory, so the benefit cannot be explained by retrieving information from LTM; rather, greater familiarity with the individual pieces, analogous to HF words, would make it easier to bind them to each other, which results in the observed advantage in immediate memory for children. Given that the adults should have stronger chunks for the pieces, the explanation of superior performance for children with random configurations likely comes from children having more "raw" WM resources.

Other studies showing that young children perform less well than older children even when the stimuli are well learned (e.g., Cowan et al., 2015) are called into question by the results described above (i.e., that familiarity advantages remain even for stimuli that we assume have "ample" exposure – chess pieces were well known to all participants). Furthermore, it is hard to accept Cowan et al.'s (2015) result as unambiguous evidence against the current proposal, since all age groups in that study were presented with WM arrays of three items for only 750 ms. Given that the speed of both voluntary and reflexive eye-movements increases dramatically until 25 years of age (Fischer, Biscaldi, & Gezeck, 1997), it is possible that younger children simply could not

encode as many items in the same time period. Finally, research on developmental reversals has suggested that, rather than having less attentional resources, younger children distribute their attention differently – globally to all features of the stimuli, rather than focusing on task-relevant information only (Deng & Sloutsky, 2016; Plebanek & Sloutsky, 2017). As a result, relative to adults, children exhibit better memory for task-irrelevant information in change-detection and visual search tasks (Plebanek & Sloutsky, 2017) and they are also better at learning accidental category features (Deng & Sloutsky, 2016).

There also have been theoretical arguments that seem somewhat consistent with our own. For example, Munakata (2001) argued that graded representations can explain phenomena that have previously been taken as evidence for discrete developmental stages. She aptly points out that some tasks require stronger representations than others. Quite recently, Nassar, Helmers, and Frank (2018) have also put forward a theory that posits a knowledge/WM trade-off. It is important to note that the veracity of the theory presented here does not rest on the claim that children might have just as much (or more) WM resources as adults. We find this to be an interesting speculation suggested by the theory and that it is not inconsistent with current research. A weaker version of this claim suggests that the developmental increase in WM capacity has been overstated with current measurement methods.

6. *Could the puzzle of second language learning be explained by the WM thesis?*

This research project was started over a decade ago and was motivated by an important puzzle in human cognition: *why is second language learning so much easier for children than adults?* While we do not discount prior explanations such as total time on task (Johnson & Newport, 1989) nor a "critical period" for fluent language acquisition (Johnson & Newport, 1989; Lenneberg, 1967), our speculative explanation for this phenomenon is not based on either, but rather follows from the other axioms we have put forward. First, children tend to process novel languages from the bottom-up – they tend to hear the phonemes and language structure before they feel compelled to speak in the new language or try to parse language word by word.

If children indeed have more WM and they are not initially trying to map the sounds onto structures in their LTM, they can build up lower level chunks that are based on phonemes and syntactic structures that are different from their first language. Young children, with a lot of exposure to speech, without the burden on WM of trying to produce complex thoughts, can strengthen the low level phonemic and syntactic structures. In contrast, adults tend to hear much more complex sentences that they try to parse and map onto their native tongue. Given our contention that adults do not have a larger WM pool, by attempting to parse the new language, they are exhausting their resources and are much more prone to *partial match* the input with existing linguistic structures. Adults have larger and stronger chunks that make partial matching easier to do. Given the greater complexity and demands of the input they receive, partial matching is of greater necessity. This could explain why adults tend to reinforce the wrong pronunciation and fail to encode the nuances of the new language. They are forced to encode too much information in a short period of time and thus rely on partial matching to get the job done.

C. Relationship to other models

SAC bears similarity to other memory models. Most notably, it evolved from the ACT-R cognitive architecture (Anderson et al., 2004). With ACT-R, it shares the assumption that memory consists of a network of nodes that differ in strength, and that the more outgoing connections there are from each node, the less activation spreads along each of those to nearby nodes. It differs from ACT-R in multiple ways; notably, the notion of a limited resource in ACT-R is implemented as a cap on the total activation that can spread in the system, rather than as a resource that is depleted for each memory operation and recovers over time. Furthermore, SAC stands with other models in the dual-processing framework (Diana et al, 2006; Yonelinas, 2002), by assuming that recognition decisions can be based either on a recollection process derived from episodic nodes, or a familiarity process derived from semantic nodes.

1. Word frequency effects

We are not aware of models that have attempted to account comprehensively for the whole range of word-frequency effects across tasks and manipulations presented here; nevertheless, word-frequency effects in each paradigm have been an impetus for empirical research and model development (for an earlier review, see Diana & Reder, 2006, or Reder et al., 2007). For single word recognition, SAC's assumption that HF words have been experienced in more diverse contexts and thus less activation spreads to any one of them during recognition (Reder et al., 2000) is similar to a mechanism in the BCDMEM model (Dennis & Humpreys, 2001). In BCDMEM recognition depends on retrieving the correct context vector, the probability of which decreases when there are more context vectors stored. This is a strong divergence from models like REM (Shiffrin & Steyvers, 1997), according to which HF items have more features in common with each other, which makes them more confusable.

If the LF recognition advantage is due to less contextual competition, as we suggest, then how would we explain Nelson and Shiffrin's (2013) results, who found in their training study that recognition performance decreased with increasing exposure frequency for previously unknown Chinese characters? Several things make it difficult to interpret their findings. First, recognition for characters that were never trained was actually intermediate between recognition for the characters in the LF and HF exposure conditions. We found the same in Reder's prior (unpublished) work, before the stimuli were carefully designed to avoid partial matching (see the section on Partial Matching). Thus, we do not know how effective Nelson & Shiffrin's (2013) training was. Second, even though Nelson and Shiffrin (2013), Reder et al. (2016) and Shen et al. (2018) used very similar training procedures (e.g., Chinese characters exposed during visual search training over multiple hour-long sessions), there is one crucial difference between them. Despite Nelson and Shiffrin's claim that contextual competition should not differ among items trained at different frequencies, each target in their study was presented with different distractors on each repetition. Since their distractors were drawn at random from each frequency class, over time, HF characters would become not only more frequently exposed, but would also become associated

with a greater contextual fan of distractors. In contrast, both Reder et al. (2016) and Shen et al. (2018) grouped characters in sets of four, and three of the characters from the set would be shown in the visual display (including or not including the target, depending on whether it was a present or absent trial). Thus, HF and LF characters did not differ much in the variability of their encoding contexts. This possibility is consistent with earlier work by Reder et al. (2002), where we independently manipulated exposure frequency for pseudowords, and the contextual fan within each frequency class.

2. Mixed-list paradox

Watkins et al. (2000) suggest that the mixed-list paradox is the result of participant controlled strategies in that participants pay more attention to LF words and ignore HF words in mixed lists. This cannot be the case, because the mixed-list paradox appears with incidental learning in both free recall (Dewhurst, Brandt, & Sharp, 2004) and serial recall (Morin et al., 2006). Another possible explanation comes from Hulme et al. (2003), who proposed that in pure HF lists, it is easier to form inter-item associations between HF words, due to their higher co-occurrence in language. These inter-item associations help retrieval, but they are removed in mixed-lists. However; this account cannot explain why LF items benefit in mixed-lists; it also cannot explain why performance is better on lists where the first half contains HF items, and the second contains LF items, compared to lists in which the first half contains LF items, and the second contains HF items (Watkins, 1977). Finally, it cannot explain why the discrimination between LF targets and foils improves as the proportion of HF words on the list increases (Malmberg & Murnane, 2002).

SAC's explanation of the mixed-list paradox is similar to the item-order hypothesis (Serra & Nairne, 1993) according to which low-frequency items require more resources for processing, leaving fewer resources for encoding the order of items. Serra & Nairne (1993) argued that in free recall tasks participants store information about the order in which items appear during study (i.e. inter-item associations), and then they use that order to guide recall despite the task being free recall. That is, even though they are not required to recall items in sequence, there is a strong tendency to do so which results in contiguity effects (for a review, see Healey et al., 2018). Thus, if there were not enough resources to encode order information, recall performance would be diminished in mixed-lists (Serra & Nairne, 1993). The difference between their verbal account and SAC is that SAC proposes a more general resource depletion mechanism – depleting more resources for LF items leaves fewer resources for processing subsequent items rather than for processing the order of items per se. This difference allows SAC to explain why the mixed-list paradox appears in recognition (Malmberg & Murnane, 2002) or source memory (Popov et al., 2019), wherein memory decisions for individual items do not benefit from remembering the order in which items appeared during study.

3. *Preceding item effects*

One family of models that explicitly deals with order effects includes TCM (Howard & Kahana, 2002) and CMR (Sederberg et al., 2008). In these models, it is assumed that a short-term memory buffer consists of a temporal context vector, which is updated with the representation of each studied item. This context vector is later used to cue memory for retrieval, and it produces various other order effects, such as the contiguity lag effect (Healey et al., 2018). It is not immediately clear what TCM/CMR would predict with respect to the findings we reported in the "Preceding item strength" section. These would be worth investigating in the future.

D. Model limitations and future directions

Despite SAC's success in modeling frequency effects in memory, the model was originally developed to capture feeling of knowing's role in strategy choice for question answering (e.g., Reder & Schunn, 1996; Schunn et al., 1997) and recognition memory performance (Reder et al., 2000). As a result, our extensions to recall tasks could be considered lacking in some respects. For example, while our serial recall model does a good job capturing the interaction between word-frequency and serial position, it does not reproduce the one item recency effect, which has been attributed to access from a WM buffer (Anderson et al., 1998). Furthermore, we have made no attempt to model the full specter of contiguity effects in free recall, which have been a crucial benchmark for models of free recall (Sederberg et al., 2008).

1. *Parameter variability*

Looking at Table 2, one could wonder why parameters such as W and w_r, which are related to WM capacity, are so variable across simulations. A number of factors that contribute to this variability. First, note that W is fixed for half of the experiments – we estimated its best fitting value for Reder et al. (2000) as 3 and we fixed this value unless the fit was not satisfactory. The experiments in which we allowed W to vary were the four serial recall experiments by Hulme et al. (1997; 2003). These were the only experiments in which we modeled the serial position curve in addition to the frequency effect and it was the serial position curve fit that was the primary driver of the change in W. Second, the variability in W and w_r is caused by the combination of two other factors – since the other learning parameters are fixed, all of the natural variability due to differences in the experimental paradigms must be captured by the two free WM parameters. We could potentially allow all parameters to vary, as is a common practice in memory modeling, but we prefer to fix the learning parameters given how well the model has performed in the past with those values.

2. *Additional predictions*

Our model makes multiple novel predictions that have yet to be tested. First, we expect that the novel effects of preceding item frequency presented in this paper should depend on the presentation rate during study. Since resources recover over time, the frequency of the preceding item should

have a smaller impact on memory for the current item with slower presentation rates and for people with larger WM capacity (similar to findings concerning the primacy effect that we reviewed above). Finally, the model presented here also captures primacy effects, directed forgetting effects (Popov et al., in press), various aspects of the spacing effect, and the null and the negative list-strength effect (Ratcliff, Clark, & Shiffrin, 1990; Wilson & Criss, 2017). These extensions could be explored in future work.

3. Temporal isolation effects

One prediction that could be challenged by existing data is that because resources recover over time, increasing the temporal gap before a study item should increase memory for that item. Most serial recall studies have not found evidence for this prediction (Brown & Lewandowsky, 2005; Lewandowsky, Brown, Wright, & Nimmo, 2006; Peteranderl & Oberauer, 2018). Nevertheless, such temporal isolation effects do occur in free recall tasks (Brown, Morin, & Lewandowsky, 2006), probed recognition tasks (Morin, Brown, & Lewandowsky, 2010), running memory span tasks (Geiger & Lewandowsky, 2008), unconstrained order reconstruction tasks (Lewandowsky, Nimmo, & Brown, 2008), and in one case, even in a forward serial recall task (Morin, Brown, & Lewandowsky, 2010)[24].

How do we reconcile the fact that the size of the pre-item interval sometimes affects performance and sometimes does not? In the case of serial recall tasks, there are two critical differences between the studies that failed to find pre-item interval effects (Brown & Lewandowsky, 2005; Lewandowsky et al., 2006; Peteranderl & Oberauer, 2018)[25] and the Morin et al. (2010) study that found the effect.

First, even though all studies attempted to prevent rehearsal during the inter-stimulus-interval (ISI), in all earlier studies participants had to continuously repeat the same word during encoding (i.e., articulatory suppression), while in Morin et al. (2010) a sequence of random digits was presented one at a time during the ISI, and participants had to read them out loud (i.e., continual distraction). Morin et al. (2010) argued that verbalizing these rapidly changing distractors is more

---

[24] Of these studies, only Brown et al. (2006) and Morin et al. (2010) have reported the effects of the pre-item and post-item interval separately rather than combined but both found that as the pre-item interval increases, free and serial recall performance increases, consistent with the predictions of the resource depletion account. Lewandowsky et al (2008) noted that they do not report the independent effects of these variables, because they have typically found that both effects are present or absent at the same time (S. Lewandowsky, personal communication, April 8, 2019)

[25] It should also be noted that a null effect is not necessarily evidence for the absence of an effect. Of these studies, only Peteranderel & Oberauer (2018) used Bayesian statistics, which are able to quantify evidence supporting the absence of an effect. The BF supporting the null hypothesis that the pre-item interval does not affect performance was 1.25, which is considered ambiguous and thus we cannot conclude anything about the presence or absence of the effect.

likely to successfully prevent the processing of the preceding items during the ISI. Thus, studies in which articulatory suppression was used did not find an effect of the ***pre-item interval*** possibly because resources were still being spent on processing the preceding item. If that is the case, we would expect to see that the size of the ***post-item interval*** predicts performance. The evidence for this is, again, mixed, with some studies finding no effect of the post-item interval (Nimmo & Lewandowsky, 2006; Parmentier, King, & Dennis, 2006), while others do find such an effect (Lewandowsky et al., 2006; Peteranderl & Oberauer, 2018)[26].

The second difference concerns whether the stimulus set forms a closed or an open pool of study items. When a closed pool of items is used, the same small number of digits or letters is studied on each list (Brown & Lewandowsky, 2005; Lewandowsky et al., 2006). It is possible that using a closed item pool can mask any potential temporal gap effects, because fewer resources are spent overall to encode these strong and often repeated items. Morin et al. (2010) and Brown et al. (2006), used unique words on each trial, thus requiring more resources for processing, which lead the pre-item interval to positively predicts performance in serial recall (Morin et al., 2010).

4. Performance decline over successive trials

The resource depletion and recovery assumptions predict that performance should gradually decline over successive trials of a memory test, and it should bounce back after a break. While this effect has not been observed (e.g. Hitch, Flude, & Burgess, 2009; Page, Cumming, Norris, McNeil, & Hitch, 2013), it should be noted that whether such an effect would occur depends on the presentation rate and the inter-trial-interval. The median time for full resource recovery in all of our simulations was 4.83 seconds, meaning that this prediction would be observed only with very short inter-trial intervals. That was not the case in the above mentioned studies (for example, Hitch et al., 2009, had a 12 second break between trials). Furthermore, if there is a full recall test in-between trials, that would allow still more resource recovery, thereby masking these potential effects. Our preliminary data show that when the inter-trial interval is kept short (750 ms), and participants have to respond to a single probe item between trials, we observe exactly the predicted decline over successive trials followed by a bounce back after a break, even when there is no general fatigue over time (Popov, So & Reder, 2019).

VI. Epilogue: The concept of resources as an explanation

To conclude, it might be good to ask the following question – how useful is the concept of resources? One could argue that resource-based explanations are circular – worse performance is attributed to having fewer resources, which on the face of it does not seem to be a satisfactory

[26] Some of these authors attribute the effect of the post-item interval to grouping (Farrell, Wise, & Lelièvre, 2011; Lewandowsky, Brown, Wright, & Nimmo, 2006) rather than to increased time for encoding/consolidation (Peteranderl & Oberauer, 2018).

explanation. While we do agree that often resource-based verbal theories can suffer from circularity, we believe that this is not the case here. Our mechanistic model includes a precise formulation of resources, which allows us to make quantifiable predictions that go well beyond the generic less-resources-worse-performance claim. Global and local list-composition effects are one such example. The current theory posits that resources do not recover immediately after use, but rather recovers slowly over time, which led to the prediction that word frequency and repetition effects would depend on the specific sequence of trials, and on the presentation speed.

One aspect that would require further work is specifying *the form* of the resource recovery function. We chose a linear recovery rate, which worked well for the studies modeled here; nevertheless, there is no a priori reason for this. Within most simulations used in the paper, WM recovers fully within several seconds. Yet, we all experience growing fatigue over the course of the day and WM and LTM performance varies as a function of time of day (for a review, see Schmidt, Collette, Cajochen, & Peigneux, 2007). Thus, it is possible that there is a component of resource recovery that operates on a longer-time scale than the one presented here, just as there are multiple fuel sources and corresponding recovery rates for muscle contraction, which is the analogy we used in the introduction to this paper. Additional work is necessary to identify such components and to provide more data for constraining the possible functional forms of the resource recovery rate/s.

Finally, the current theory provides a mechanistic account for why additional learning benefits weaker memory traces more. When items are restudied, stronger representations in the model are strengthened to a lesser degree than weaker representations, due to the delta learning function. The explanation for why this function is suitable comes from a rational perspective – every strengthening depletes resources proportional to the amount of strengthening. Since resources are limited, it is most optimal to spend less for processing stronger items, because they have a lower probability of being forgotten, and to reserve those resources for processes that need them more. We argue that, similar to partial matching, this learning function reflects an adaptation to the challenge that there is typically much more information demanding our attention than our limited resources allow us to process.

Acknowledgements

We are grateful to Rachel Diana who initiated this line of work in her dissertation with L. Reder. We are also grateful to Xiaonan Liu who found the paper and researchers that enabled us to examine similarity effects among Chinese characters. We are also grateful to Klaus Oberauer, Ed Awh, Nelson Cowan and Jason Sandor for fruitful discussions concerning many of the phenomena reported, and for suggesting some plausible alternative explanations. We thank John Anderson, Marc Coutanche, Markus Ostarek and Matthew So for commenting on previous drafts of the paper. Finally, we are grateful to the researchers who shared their data for the novel re-analyses reported in section IV.

References

Ahmad, F. N., Fernandes, M., & Hockley, W. E. (2015). Improving associative memory in older adults with unitization. *Aging, Neuropsychology, and Cognition*, *22*(4), 452–472.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97–123.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. L. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

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.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*(3), 221–256.

Anderson, N. D., Iidaka, T., Cabeza, R., Kapur, S., McIntosh, A. R., & Craik, F. I. (2000). The effects of divided attention on encoding-and retrieval-related brain activity: A PET study of younger and older adults. *Journal of Cognitive Neuroscience*, *12*(5), 775–792.

Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L., Jennett, S., … Seidler, R. D. (2012). The effects of working memory resource depletion and training on sensorimotor adaptation. *Behavioural Brain Research*, *228*(1), 107–115.

Anguera, J. A., Reuter-Lorenz, P. A., Willingham, D. T., & Seidler, R. D. (2010). Contributions of spatial working memory to visuomotor learning. *Journal of Cognitive Neuroscience*, *22*(9), 1917–1930.

Aue, W. R., Criss, A. H., & Novak, M. D. (2017). Evaluating mechanisms of proactive facilitation in cued recall. *Journal of Memory and Language*, *94*, 103–118.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 576.

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language*, *46*(1), 199–226.

Bartsch, L. M., Loaiza, V. M., & Oberauer, K. (2019). Does limited working memory capacity underlie age differences in associative long-term memory?. *Psychology and aging*, *34*(2), 268.

Baumeister, R. F., Bratslavsky, E., & Muraven, M. (2018). Ego depletion: Is the active self a limited resource?. In *Self-Regulation and Self-Control* (pp. 24-52). Routledge.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, *2*, 35-67.

Brockmole, J. R., & Logie, R. H. (2013). Age-Related Change in Visual Working Memory: A Study of 55,753 Participants Aged 8–75. *Frontiers in Psychology*, *4*.

Brown, G. D., & Lewandowsky, S. (2005). Serial recall and presentation schedule: A micro-analysis of local distinctiveness. *Memory*, *13*(3–4), 283–292.

Brown, G. D., Morin, C., & Lewandowsky, S. (2006). Evidence for time-based models of free recall. *Psychonomic bulletin & review*, *13*(4), 717-723.

Buchler, N. E. G., & Reder, L. M. (2007). Modeling age-related memory deficits: A two-parameter solution. *Psychology and Aging*, *22*(1), 104–121.

Buchler, N. G., Faunce, P., Light, L. L., Gottfredson, N., & Reder, L. M. (2011). Effects of Repetition on Associative Recognition in Young and Older Adults: Item and Associative Strengthening. *Psychology and Aging*, *26*(1), 111–126.

Buchler, N. G., Light, L. L., & Reder, L. M. (2008). Memory for items and associations: Distinct representations and processes in associative recognition. *Journal of Memory and Language*, *59*(2), 183–199.

Cabeza, R., Mangels, J., Nyberg, L., Habib, R., Houle, S., McIntosh, A. R., & Tulving, E. (1997). Brain regions differentially involved in remembering what and when: a PET study. *Neuron*, *19*(4), 863–870.

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language*, *61*(3), 457–469.

Caplan, J. B., Madan, C. R., & Bedwell, D. J. (2015). Item-properties may influence item-item associations in serial recall. *Psychonomic Bulletin & Review*, *22*(2), 483–491.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*(2), 231–248.

Chalfonte, B. l., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory & Cognition*, *24*(4), 403-416. doi:10.3758/bf03200930

Chalmers, K. A., & Humphreys, M. S. (2003). Experimental manipulation of prior experience: Effects on item and associative recognition. *Memory*, *11*(3), 233–246.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chee, M. W., Westphal, C., Goh, J., Graham, S., & Song, A. W. (2003). Word frequency and subsequent memory effects studied using event-related fMRI. *NeuroImage*, *20*(2), 1042-1051.

Chi, M. T. (1978). Knowledge structures and memory development. *Children's Thinking: What Develops*, *1*, 75–96.

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*(3), 257–273.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition*, *20*(3), 231–243.

Clark, S. E., & Burchett, R. E. R. (1994). Word frequency and list composition effects in associative recognition and recall. *Memory & Cognition*, *22*(1), 55–62.

Clark, S. E., & Shiffrin, R. M. (1992). Cuing effects and associative information in recognition memory. *Memory & Cognition*, *20*(5), 580–598.

Coltheart, V., Mondy, S., Dux, P. E., & Stephenson, L. (2004). Effects of orthographic and phonological word length on memory for lists shown at RSVP and STM rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 815.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114-185.

Cowan, N., Naveh-Benjamin, M., Kilb, A., & Saults, J. S. (2006). Life-span development of visual working memory: When is feature binding difficult? *Developmental Psychology*, *42*(6), 1089–1102.

Cowan, N., Ricker, T. J., Clark, K. M., Hinrichs, G. A., & Glass, B. A. (2015). Knowledge cannot explain the developmental growth of working memory capacity. *Developmental Science*, *18*(1), 132–145.

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, *147*(4), 545.

Craik, F. I., & Byrd, M. (1982). Aging and cognitive deficits. In *Aging and cognitive processes* (pp. 191–211). Springer.

Craik, F. I., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of verbal learning and verbal behavior*, *12*(6), 599-607.

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language*, *55*(4), 447–460.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language*, *64*(2), 119–132.

Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, *25*(3), 315–353.

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports*, *7*, 337–344.

DeLosh, E., & Mcdaniel, M. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1136–1146

Deng, W. S., & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, *91*, 24–62.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452.

Dewhurst, S. A., Brandt, K. R., & Sharp, M. S. (2004). Intention to learn influences the word frequency effect in recall but not in recognition memory. *Memory & cognition*, *32*(8), 1316-1325.

DeWitt, M. R., Knight, J. B., Hicks, J. L., & Ball, B. H. (2012). The effects of prior knowledge on the encoding of episodic contextual details. *Psychonomic Bulletin & Review*, *19*(2), 251–257.

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory & Cognition*, *33*(7), 1289–1302.

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 805.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, *13*(1), 1–21.

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*(4), 873–902.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like Models of Recognition and Learning*. *Cognitive Science*, *8*(4), 305–336.

Fischer, B., Biscaldi, M., & Gezeck, S. (1997). On the development of voluntary and reflexive components in human saccade generation. *Brain Research*, *754*(1), 285–297.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, *134*(4), 521.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental psychology*, *40*(2), 177.

Geiger, S. M., & Lewandowsky, S. (2008). Temporal isolation does not facilitate forward serial recall—or does it?. *Memory & Cognition*, *36*(5), 957-967.

Gilbert, A. C., Boucher, V. J., & Jemel, B. (2014). Perceptual chunking and its effect on memory in speech processing: ERP and behavioral evidence. *Frontiers in Psychology*, *5*, 220.

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

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 5.

Glenberg, A., Smith, S. M., & Green, C. (1977). Type I rehearsal: Maintenance and more. *Journal of Verbal Learning and Verbal Behavior*, *16*(3), 339-352.

Gobet, F. (1998). Expert memory: a comparison of four theories. *Cognition*, *66*(2), 115–152.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: evidence for the magical number four ... or is it two? *Memory (Hove, England)*, *12*(6), 732–747.

Gobet, F., & Simon, H. A. (1996). Recall of rapidly presented random chess positions is a function of skill. *Psychonomic Bulletin & Review*, *3*(2), 159–163.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Grady, C. L., McIntosh, A. R., Horwitz, B., Maisog, J. M., Ungerleider, L. G., Mentis, M. J., … Haxby, J. V. (1995). Age-related reductions in human recognition memory due to impaired encoding. *Science*, *269*(5221), 218–221.

Gregg, V. H., Montgomery, D. C., & Castaño, D. (1980). Recall of common and uncommon words from pure and mixed lists. *Journal of Verbal Learning and Verbal Behavior*, *19*(2), 240–245.

Healey, M. K., Long, N. M., & Kahana, M. J. (2018). Contiguity in episodic memory. *Psychonomic bulletin & review*, 1-22.

Healey, M. K., & Kahana, M. J. (2016). A four–component model of age–related memory change. *Psychological Review*, 123(1), 23-69

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

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 201–205.

Hitch, G. J., Flude, B., & Burgess, N. (2009). Slave to the rhythm: Experimental tests of a model for verbal short-term memory and long-term sequence learning. *Journal of Memory and Language*, *61*(1), 97-111.

Hockley, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition*, *22*(6), 713–722.

Horner, A. J., & Henson, R. N. (2008). Priming, response learning and repetition suppression. *Neuropsychologia*, *46*(7), 1979-1991.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269-299.

Huber, D. E., Clark, T. F., Curran, T., & Winkielman, P. (2008). Effects of repetition priming on recognition memory: testing a perceptual fluency-disfluency model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1305.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology-Learning Memory and Cognition*, *23*(5), 1217–1232.

Hulme, C., Stuart, G., Brown, G. D., & Morin, C. (2003). High-and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory and Language*, *49*(4), 500-518.

Humphreys, M. S., Maguire, A. M., McFarlane, K. A., Burt, J. S., Bolland, S. W., Murray, K. L., & Dunn, R. (2010). Using maintenance rehearsal to explore recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 147–159.

Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 21.

Jennings, J. M., & Jacoby, L. L. (1993). Automatic versus intentional uses of memory: Aging, attention, and control. *Psychology and Aging*, *8*(2), 283.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99.

Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1534.

Kamas, E. N., Reder, I. M., & Ayers, M. S. (1996). Partial matching in the Moses illusion: Response bias not sensitivity. *Memory & Cognition*, *24*(6), 687–699.

Lenneberg, E. H. (1967). The Biological Foundations of Language. *Hospital Practice*, *2*(12), 59–67.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*(1), 25–57.

Lewandowsky, S., Brown, G. D., Wright, T., & Nimmo, L. M. (2006). Timeless memory: Evidence against temporal distinctiveness models of short-term memory for serial order. *Journal of Memory and Language*, *54*(1), 20–38.

Lewandowsky, S., Nimmo, L. M., & Brown, G. D. (2008). When temporal isolation benefits memory for serial order. *Journal of Memory and Language*, *58*(2), 415-428.

Light, L. L., Patterson, M. M., Chung, C., & Healy, M. R. (2004). Effects of repetition and response deadline on associative recognition in young and older adults. *Memory & Cognition*, *32*(7), 1182–1193. https://doi.org/10.3758/BF03196891

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological review*, *95*(4), 492.

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency effect in memory for mixed frequency lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(6), 1943–1946.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Cognitive Systems Research*, *1*(2), 99–118.

Lovett, M. C., Reder, L. M., & Lebiere, C. (1999). Modeling working memory in a unified architecture: An ACT-R perspective. In Miyake, A. and Shah, P. (Eds). *Models of Working Memory*. Cambridge University Press, 135-182.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 132.

Madan, C. R., Glaholt, M. G., & Caplan, J. B. (2010). The influence of item properties on association-memory. *Journal of Memory and Language*, *63*(1), 46–63.

Maddox, W. T., & Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 539.

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 616–630.

Malmberg, K. J., & Nelson, T. O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition*, *31*(1), 35–43.

Mandler, G., Goodman, G. O., & Wilkes-Gibbs, D. L. (1982). The word-frequency paradox in recognition. *Memory & Cognition*, *10*(1), 33–42.

Marevic, I., Arnold, N. R., & Rummel, J. (2017). Item-Method Directed Forgetting and Working Memory Capacity: A Hierarchical Multinomial Modeling Approach. *The Quarterly Journal of Experimental Psychology*, 1–34.

Marshall, P. H., & Werder, P. R. (1972). The effects of the elimination of rehearsal on primacy and recency. *Journal of Verbal Learning and Verbal Behavior*, *11*(5), 649-653.

Martin, J. G. (1964). Associative strength and word frequency in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, *3*(4), 317-320.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724–760.

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

Miller, L. M., & Roodenrys, S. (2012). Serial recall, word frequency, and mixed lists: The influence of item arrangement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1731.

Morin, C., Brown, G. D., & Lewandowsky, S. (2010). Temporal isolation effects in recognition and serial recall. *Memory & Cognition*, *38*(7), 849-859.

Morin, C., Poirier, M., Fortin, C., & Hulme, C. (2006). Word frequency and the mixed-list paradox in immediate and delayed serial recall. *Psychonomic Bulletin & Review*, *13*(4), 724–729.

Munakata, Y. (2001). Graded representations in behavioral dissociations. *Trends in cognitive sciences*, *5*(7), 309-315.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 855.

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological review*, *125*(4), 486.

Neath, I., & Crowder, R. G. (1996). Distinctiveness and very short-term serial position effects. *Memory*, *4*(3), 225-242.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394.

Nickerson, R. S., & Adams, M. J. (1979). Long-term memory for a common object. *Cognitive Psychology*, *11*(3), 287–307.

Nimmo, L. M., & Roodenrys, S. (2002). Syllable frequency effects on phonological short-term memory tasks. *Applied Psycholinguistics*, *23*(04), 643–659.

Oates, J. M., Reder, L. M., Cook, S. P., & Faunce, P. (2015). Spurious Recollection from a Dual-Process Framework. *Cognitive Modeling in Perception and Memory: A Festschrift for Richard M. Shiffrin*, 145-161.

Oberauer, K. (2003). Understanding serial position curves in short-term recognition and recall. *Journal of Memory and Language*, *49*(4), 469–483.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological review*, *122*(2), 260.

Osth, A. F., Fox, J., McKague, M., Heathcote, A., & Dennis, S. (2018). The list strength effect in source memory: Data and a global matching model. Journal of Memory and Language, 103, 91-113.

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.

Ozubko, J. D., & Joordens, S. (2007). The mixed truth about frequency effects on free recall: Effects of study list composition. *Psychonomic Bulletin & Review*, *14*(5), 871–876.

Page, M. P., Cumming, N., Norris, D., McNeil, A. M., & Hitch, G. J. (2013). Repetition-spacing and item-overlap effects in the Hebb repetition task. *Journal of Memory and Language*, *69*(4), 506-526.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*(4), 441–474.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 324.

Palmeri, T. J., & Tarr, M. (2008). Visual object perception and long-term memory. *Visual memory*, 163-207.

Palmeri, T. J., Wong, A. C., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in cognitive sciences*, *8*(8), 378-386.

Parmentier, F. B., King, S., & Dennis, I. (2006). Local temporal distinctiveness does not benefit auditory verbal and spatial serial recall. *Psychonomic Bulletin & Review*, *13*(3), 458-465.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*(6), 528-530.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. *Cognitive Science*, *29*(4), 559–586.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.

Peteranderl, S., & Oberauer, K. (2018). Serial recall of colors: Two models of memory for serial order applied to continuous visual stimuli. *Memory & cognition*, *46*(1), 1-16.

Peterson, D. J., & Naveh-Benjamin, M. (2016). The role of aging in intra-item and item-context binding processes in visual working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(11), 1713–1730.

Peterson, D. J., & Naveh-Benjamin, M. (2017). The role of attention in item-item binding in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1403.

Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of Selective Attention: When Children Notice What Adults Miss. *Psychological Science*, *28*(6), 723–732.

Popov, V., Marevic, I., Rummel, J., & Reder, L. (in press). Forgetting is a Feature, not a Bug: Intentionally Forgetting Some Things Helps Us Remember Others by Freeing up Working Memory Resources. *Psychological Science*

Popov, V., So, M., & Reder, L. (2019, May 23). Word frequency affects binding probability not memory precision. https://doi.org/10.31234/osf.io/deyjm

Potter, M. C., Staub, A., Rado, J., & O'connor, D. H. (2002). Recognition memory for briefly presented pictures: the time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(5), 1163.

Rao, K. V., & Proctor, R. W. (1984). Study-phase processing and the word frequency effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 386.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology*, *19*(1), 90–138.

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language*, *30*(4), 385–406.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 435.

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. Retrieved from http://d-scholarship.pitt.edu/22852/1/licence.txt

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A reexamination of stimulus-frequency effects in recognition: Two mirrors for low- and high-frequency

pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 138–152.

Reder, L. M., Liu, X. L., Keinath, A., & Popov, V. (2016). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic Bulletin & Review*, *23*(1), 271–277.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 294.

Reder, L. M., Park, H., & Kieffaber, P.D. (2009). Memory systems do not divide on consciousness: Reinterpreting memory in terms of activation and binding. *Psychological Bulletin, 135*(1), 23-49.

Reder, L. M., Paynter, C., Diana, R. A., Ngiam, J., & Dickison, D. (2007). Experience is a Double-Edged Sword: A Computational Model of The Encoding/Retrieval Trade-Off With Familiarity. In *Psychology of Learning and Motivation* (Vol. 48, pp. 271–312). Elsevier.

Reder, L. M., Victoria, L. W., Manelis, A., Oates, J. M., Dutcher, J. M., Bates, J. T., … Gyulai, F. (2013). Why It's Easier to Remember Seeing a Face We Already Know Than One We Don't Preexisting Memory Representations Facilitate Memory Formation. *Psychological Science*, *24*(3), 363–372.

Reed, A. V. (1973). Speed-Accuracy Trade-Off in Recognition Memory. *Science*, *181*(4099), 574–576. Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*(4099), 574-576.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Roodenrys, S., Hulme, C., & Brown, G. (1993). The development of short-term memory span: Separable effects of speech rate and long-term memory. *Journal of Experimental Child Psychology*, *56*(3), 431–442.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of experimental psychology*, *89*(1), 63.

Sala, G., & Gobet, F. (2017). Experts' memory superiority for domain-specific random material generalizes across fields of expertise: A meta-analysis. *Memory & Cognition*, *45*(2), 183–193.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, *27*(5), 763–776. https://doi.org/10.1037/0012-1649.27.5.763

Schmidt, C., Collette, F., Cajochen, C., & Peigneux, P. (2007). A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology*, *24*(7), 755–789.

Schneider, D. W., & Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology*, *64*(3), 127–160.

Schneider, W., Gruber, H., Gold, A., & Opwis, K. (1993). Chess Expertise and Memory for Chess Positions in Children and Adults. *Journal of Experimental Child Psychology*, *56*(3), 328–349.

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, *9*(4), 211–212.

Schulman, A. I. (1976). Memory for rare words previously rated for familiarity. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(3), 301.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 3.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893–912.

Serra, M., & Nairne, J. S. (1993). Design controversies and the generation effect: Support for an item-order hypothesis. *Memory & Cognition*, *21*(1), 34-40.

Shallice, T., Fletcher, P., Frith, C. D., Grasby, P., Frackowiak, R. S. J., & Dolan, R. J. (1994). Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature*, *368*(6472), 633–635

Shen, Z., Popov, V., Delahay, A. B., & Reder, L. M. (2018). Item strength affects working memory capacity. *Memory & Cognition*.

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

Simon, H. A. (1974). How big is a chunk? *Science*, *183*(4124), 482–488.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in cognitive sciences*, *1*(7), 261-267.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176-199.

Sumby, W. H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, *1*(6), 443–450.

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

Tarr, M. (2018). *Novel Objects (Online)* (Accessed June 6, 2018). Available online at: http://wiki.cnbc.cmu.edu/Novel_Objects

Tehan, G., & Humphreys, M. S. (1988). Articulatory loop explanations of memory span and pronunciation rate correspondences: a cautionary note. *Bulletin of the Psychonomic Society*, *26*, 293–296.

Tierney, J. F., Vale, C., Riley, R., Smith, C. T., Stewart, L., Clarke, M., & Rovers, M. (2015). Individual Participant Data (IPD) Meta-analyses of Randomised Controlled Trials: Guidance on Their Use. *PLOS Medicine*, *12*(7), e1001855.

Townsend, J. T., & Ashby, F. G. (1983). Stochastic modeling of elementary psychological processes. CUP Archive.

Tulving, E., & Patkau, J. E. (1962). Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *16*(2), 83.

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*, *62*(4), 392–406.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory capacity—fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review*, *16*(5), 931–937.

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Ward, G., Woodward, G., Stevens, A., & Stinson, C. (2003). Using overt rehearsals to explain word frequency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 186.

Watkins, M. J. (1977). The intricacy of memory span. *Memory & Cognition*, *5*(5), 529–534.

Watkins, M. J., LeCompte, D. C., & Kim, K. (2000). Role of study strategy in recall of mixed lists of common and rare words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 239.

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78-88.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 681.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*(4), 1025–1054. https://doi.org/10.1037/a0020874

Xie, W., & Zhang, W. (2016). Familiarity increases the number of remembered Pokémon in visual short-term memory. *Memory & Cognition*. https://doi.org/10.3758/s13421-016-0679-7

Xie, W., & Zhang, W. (2017). Familiarity Speeds Up Visual Short-Term Memory Consolidation. *Journal of Experimental Psychology: Human Perception and Performance*.

Xing, H., Shu, H., & Li, P. (2004). The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science*, *5*(1), 1-49.

Yang, J., McCandliss, B. D., Shu, H., & Zevin, J. D. (2009). Simulating Language-specific and Language-general Effects in a Statistical Learning Model of Chinese Reading. *Journal of Memory and Language*, *61*(2), 238–257.

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517.

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of word frequency. *Bulletin of the Psychonomic Society*, *11*(6), 371–373.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.