Memory & Cognition (2021) 49:968–983
https://doi.org/10.3758/s13421-020-01117-2

How do recall requirements affect decision-making in free recall initiation? A linear ballistic accumulator approach

Adam F. Osth[1] · Aimee Reed[1] · Simon Farrell[2]

Accepted: 11 November 2020 / Published online: 2 February 2021

Abstract

Models of free recall describe free recall initiation as a decision-making process in which items compete to be retrieved. Recently, Osth and Farrell (*Psychological Review*, *126*, 578–609, 2019) applied evidence accumulation models to complete RT distributions and serial positions of participants' first recalls in free recall, which resulted in some novel conclusions about primacy and recency effects. Specifically, the results of the modeling favored an account in which primacy was due to reinstatement of the start-of-the-list, and recency was found to be exponential in shape. In this work, we examine what happens when participants are given alternative recall instructions. Prior work has demonstrated weaker primacy and greater recency when fewer items are required to report (Ward & Tan, *Memory & Cognition*, 2019), and a key question is whether this change in instructions qualitatively changes the nature of the recall process, or merely changes the parameters of the recall competition. We conducted an experiment where participants studied six- or 12-item lists and were post-cued as to whether to retrieve a single item, or as many items as possible. Subsequently, we applied LBA models with various assumptions about primacy and recency, implemented using hierarchical Bayesian techniques. While greater recency was observed when only one item was required for output, the model selection did not suggest that there were qualitative differences between the two conditions. Specifically, start-of-list reinstatement and exponential recency functions were favored in both conditions.

Keywords Free recall · Evidence accumulation models · RT distributions · Serial position effects · Linear ballistic accumulator

When given an instruction to recall as many items as possible from a list of items, how do participants initiate their recall sequence? While this is a question about memory, a number of memory models ultimately treat this as a question involving decision-making. In other words, a key component of free recall initiation is not just bringing relevant information to mind, but also a decision about which response to select for output among a set of memories. For example, in the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1981), retrieval strengths for each item are converted into sampling probabilities using Luce's choice rule (Luce, 1959). Accordingly, it isn't just memory strength that will determine recall but also the manner in which items compete for output in the decision stage. Over the past two decades, this general architecture—memory strengths being used to drive a decision-making process in which items compete for recall—has become the standard theoretical assumption for determining recall responses in free recall models (e.g., Davelaar et al., 2005; Lehman & Malmberg, 2013; Raaijmakers & Shiffrin, 1981).

A limitation of memory models based on Luce's choice rule is that they have little to say about the dynamics of recall. In particular, these models do not address the complete distributions of latency across each response, and so leave unexplained a substantial amount of data about the recall process. A modeling framework that has been successful in accounting for both choice and response time (RT) distributions in other domains is evidence accumulation models (e.g., Evans & Wagenmakers, 2020;

This work was supported by ARC Discovery Early Career award (DE170100106) awarded to Adam Osth. Data and model code can be found at https://osf.io/9vtfq/

✉ Adam F. Osth
adamosth@gmail.com

[1] University of Melbourne, Parkville, Australia

[2] University of Western Australia, Crawley, Australia

Ratcliff & Smith, 2004; Smith, 2000). In such models, a noisy process of evidence accumulation moves in the direction of two or more thresholds, each associated with a different decision outcome. Once a threshold is reached, the associated decision is made and the time taken to reach the threshold is the RT plus additional time for processes related to encoding and outputting the decision alternatives (e.g., nondecision time). Evidence accumulation models have been highly successful in accounting for both choice and RT distributions across a range of conditions in recognition memory tasks (Criss, 2010; Donkin & Nosofsky, 2012b; Fox et al., 2020; Osth et al., 2017; Ratcliff, 1978; Starns et al., 2012) but have received considerably less emphasis in free recall.

The present work focuses on applying the linear ballistic accumulator (LBA) model (Brown & Heathcote, 2008) to responses and RTs from free recall initiation, specifically to observe how the number of items required for recall affects the latent primacy mechanisms and recency functions (Osth & Farrell, 2019). Primacy and recency effects refer to recall advantages for the beginning and end-of-list items, respectively (e.g., Murdock, 1962); such advantages can also be reflected in probability of first recall (PFR) curves, where it is found that participants are most likely to initiate free recall at either the beginning or end of the list and very rarely initiate in the middle (Healey & Kahana, 2014; Howard & Kahana, 1999; Laming, 1999; Ward et al., 2010). We restrict consideration to first recalls because there are a number of constraints in modeling complete sequences that require additional assumptions. Strong sequential dependencies are present in recalled sequences—a recalled item is very likely to be followed by an item studied adjacent to that item on the study list (Healey et al., 2019; Kahana, 1996). Modeling sequential dependencies requires mechanisms such as using retrieved items (e.g., Raaijmakers & Shiffrin, 1981) or contexts (e.g., Howard & Kahana, 2002) as cues for subsequent recalls. Further mechanisms are required to account for erroneous repeated recalls (Lohnas et al., 2015) and termination of the recall sequence Harbison et al. (2009). Exploration of such mechanisms goes considerably beyond the scope and goals of the current work.

An example of an LBA model of the free recall task for a six-item list is given in Fig. 1a, with the amount of evidence for a decision depicted on the y-axis and time depicted on the x-axis. Each item is represented by an accumulator, and the accumulators race toward a common threshold (dashed line). Evidence accumulation begins at a random point for each accumulator that is sampled from a uniform distribution with height A. Accumulation is both linear and deterministic within a trial, but there is variability between accumulators across trials: on each trial an accumulator's drift rate is sampled from a normal distribution with mean v and standard deviation s. An accumulator with a higher drift rate will reach the threshold more quickly and will win the race more frequently, resulting in faster RTs and more frequent recalls. However, the accumulator with the highest drift rate is not guaranteed to win the competition—even an item with a low drift rate such as the second item *guts* can still win the competition by virtue of sampling either a high drift rate from the drift rate distribution or a high starting point from the start point distribution.

Evidence accumulation models of the free recall task

The first known application of evidence accumulation models to free recall was with variants of the temporal context model (TCM: Howard and Kahana, 2002). These variants replaced Luce's choice rule with a set of leaky competitive accumulators (LCA: Usher & McClelland, 2001) for each item (Polyn et al., 2009; Sederberg et al., 2008). In principle, these models are able to account for the full dynamics of recall as well as response probabilities. However, in practice these models have only been applied to mean RTs and not the complete distributions. This is likely due to the fact that the LCA is intractable.

Recently, Osth and Farrell (2019) jointly modeled serial position and complete RT distributions with two different evidence accumulation models: the LBA and racing diffusion model (Tillman et al., 2020) to provide novel insights into primacy and recency effects in free recall initiation. Not only did this work account for both response probabilities and latency distributions, the modeling led to several novel insights. First, while a large literature has shown advantage for power law functions to describe the shape of recency gradients (e.g., Averell & Heathcote, 2011; Donkin & Nosofsky, 2012a; Rubin & Wenzel, 1996; Wixted & Ebbesen, 1991), Osth and Farrell found consistent support for an exponential recency function. This was largely due to the fact that PFR curves are strongly peaked for the recency items and decay sharply such that mid-list items are rarely recalled. A power function, in contrast, decays gradually as lag is increased. While it may seem peculiar to find such support, models where recency effects are determined by a gradually changing context throughout presentation of the list (i.e., contextual drift) naturally produce exponential recency functions (Howard, 2004; Osth et al., 2018). In addition, the comparison between exponential and power law functions was performed at the *latent* level, similar to an approach by Donkin and Nosofsky (2012a). That is, rather than describing how manifest variables such as PFR or RT change as study-test lag is increased, the recency functions were implemented as drift rates in the evidence accumulation models.

Second, the modeling provided novel insights into the mechanism of the primacy effect. Despite decades of research on the primacy effect, there is still no consensus on

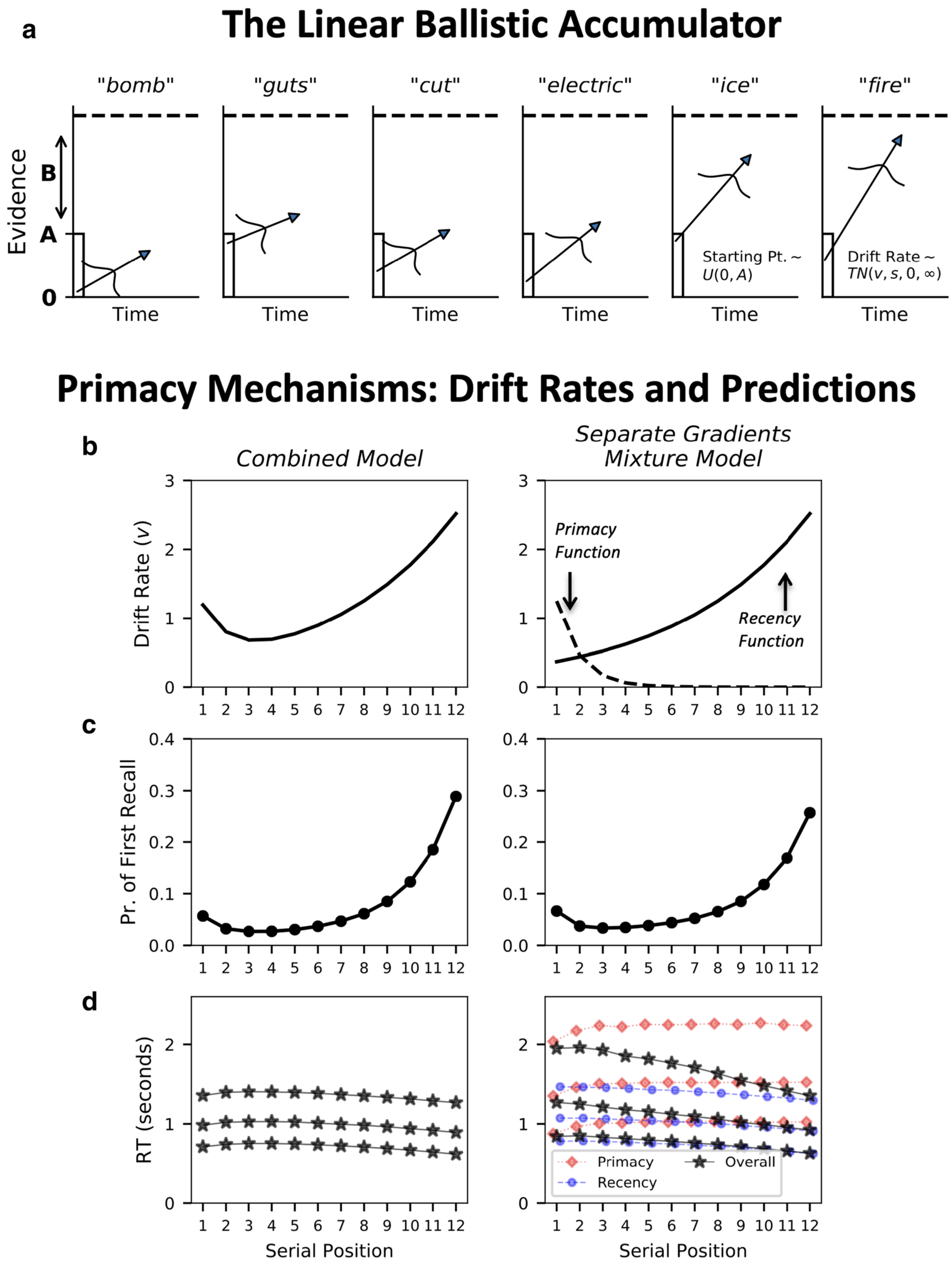

Fig. 1 Diagram of the linear ballistic accumulator (**a**) along with drift rates (**b**), predicted probability of first recall curves (**c**) and predicted response time (RT) distributions (**d**) from the LBA implementations of the primacy mechanisms, namely the combined model (*left column*) and the separate gradients mixture model (*right column*). See the text for details

its origins. A number of models postulate that the primacy and recency items directly compete to be retrieved—the primacy effect arises because early list items receive a strength boost either due to extra time in a capacity-limited buffer (Atkinson & Shiffrin, 1968; Lehman & Malmberg, 2013; Raaijmakers & Shiffrin, 1981), enhanced attention to the early list items (Serruya et al., 2014), or higher temporal distinctiveness (Brown et al., 2007). In contrast, an alternative class of models assumes that the primacy effect arises from retrieval strategies and reflects the usage of different cues for the beginning and end-of-list. In this class of models, the early list items are associated to a start-of-list marker. During recall, participants can either initiate their recalls with an end-of-list context that matches the recency items, or they can reinstate the start-of-list context to initiate from the beginning (Farrell, 2012; Metcalfe & Murdock, 1981; Morton & Polyn, 2016). Accordingly, in these models, primacy and recency items do not necessarily compete with each other to be retrieved—primacy and recency advantages arise from different cues being employed on different trials. In other words, on any one trial only the primacy items dominate the race, or only the recency items, but not both.

Osth and Farrell implemented both classes of primacy mechanisms in evidence accumulation models[1]. The class of models where primacy and recency items compete to be retrieved was implemented by assuming a single non-monotonic drift rate gradient (Fig. 1b left column). This model was referred to as the *combined* model of primacy, as a primacy gradient was directly combined with a recency gradient to form a single function. The alternative class of models where the start-of-the-list was reinstated (so that only primacy items are in the recall competition) was implemented by assuming a mixture model where on a proportion of trials p, participants initiate their recall with a primacy gradient and proportion $1 - p$ a recency gradient is employed (Fig. 1B, right column). This model was referred to as the *separate gradients* mixture model.

At first glance, both classes of models appear very similar to each other and potentially difficult to discriminate empirically. Indeed, they make very similar predictions about PFR curves (Fig. 1c): both models can capture a strong recency effect along with a weaker primacy advantage. However, the two mechanisms make different predictions about how RT distributions vary by serial position. These predictions are plotted in Fig. 1d, where RT distributions are summarized using the .1, .5, and .9 quantiles (in other words, the 10th, 50th, and 90th percentiles of the RT distribution).

Intuitively, one might expect that accumulators with higher drift rates will result in faster RTs when they are recalled. However, the simulations reveal that this is only partly true; RT distributions are predicted to be surprisingly similar across serial positions. Why is this the case? As it turns out, when there are a large number of accumulators, RTs are strongly determined by the fastest accumulator in the race (statistical facilitation, Raab, 1962). In order for a relatively weak item to be retrieved, in addition to beating out all the other accumulators, it has to also beat the accumulator with the highest drift rate. This makes it such that an accumulator with a weak drift rate, such as the item in the third serial position, will have a similar (but not identical) RT to the strongest accumulator, namely the final item.

Understanding these constraints helps us to understand the predictions of the separate gradients mixture model. Depicted are predictions from the recency race (blue) and the primacy race (red). Within each race, the predicted RTs are similar for each serial position. However, the predicted RTs from the primacy race are noticeably slower due to the fact that the strongest accumulator in the primacy gradient is weaker than the stronger accumulator in the recency gradient. When a weighted average across both races is calculated (black), RTs are predicted to vary considerably by serial position, with recalls in the first position predicted to be considerably slower than from later items despite the fact that the first item exhibits a recall advantage. Thus, the separate gradients model is capable of predicting a dissociation between accuracy and RT. Nonetheless, these predictions depend on the relative strengths of the primacy and recency items. If the strength of the primacy item was increased, the primacy race would be faster, resulting in smaller differences between the two races and a smaller difference in RTs between the first and final items.

In their model selection procedure, Osth and Farrell (2019) fit a total of 14 datasets that comprised various list lengths and recall types, including immediate, delayed, and continual distracter free recall. The modelling results strongly favored the separate gradients model, with a number of datasets showing the predicted pattern wherein the first item was recalled more slowly than succeeding items (a finding first reported by Laming, 1999). The fact that this model was supported even during delayed and continual distracter free recall is particularly constraining, as it rules out a dual store interpretation of the model. Specifically, one could interpret the recency gradient as being driven by retrieval from short-term memory (STM) and the primacy gradient as retrieval from long-term memory (LTM). However, in delayed and continual-distracter free recall datasets retrieval was delayed by a demanding distracter task that would be sufficient to clear the contents of STM, leaving only a single source (LTM) for retrieval.

[1] Osth and Farrell (2019) additionally implemented Tan and Ward (2000)'s primacy-as-recency account, in which the primacy item is functionally recent due to its rehearsal to the end of the list. However, this mechanism of primacy did not perform very well in the model selection and is not the focus of the present article.

These results support the class of models where primacy is due to a strategic reinstatement of the start-of-list context. In addition, it was an example of how a joint consideration of RT distributions and serial position was able to distinguish predictions from two mechanisms of primacy that previously mimicked each other quite strongly.

Effects of recall requirements on primacy and recency

In each of the datasets examined by Osth and Farrell (2019) participants were instructed to recall as many items as possible from the preceding study list. However, there is evidence that the number of items that participants are instructed to recall exerts a large effect on primacy and recency. Tan et al. (2016) presented participants with short lists of 4–6 items and instructed participants to recall either one, two, or three items, or a standard free recall instruction (recall as many items as possible). With standard instructions, participants exhibited primacy and recency effects of comparable magnitude. The fact that these effects were roughly of the same size is likely due to the short list lengths employed (e.g., Ward et al., 2010). However, when only one or two items were required for recall, participants showed considerably weaker primacy and stronger recency.

In that study, the instructions about recall requirements were presented in advance of the study list, which could produce differences in encoding strategies between the conditions. Specifically, when asked to recall only a single item, participants may have neglected to encode the early list items and focus more strongly on one of the recency items instead. However, Ward and Tan (2019) conducted a similar experiment where participants were *post-cued* with the recall requirements—participants were instructed as to how many items to recall after the list had already been encoded. The results replicated those of Tan et al. (2016), suggesting that the shift from primacy-to-recency as less items are required for recall is not due to encoding strategies but instead due to participants changing their decision as to which items to recall.

This article is focused on the question as to what is causing the shift from primacy-to-recency as recall requirements are lightened. Specifically, does this shift reflect a qualitative change in primacy and recency mechanisms? One possibility is that participants engage in unique strategies in the standard free recall condition, where the goal is to recall as many items as possible. In terms of primacy mechanisms, strategic reinstatement of the start-of-the-list may assist in this goal. When participants initiate with the first item, their serial position curves for the entire list strongly resemble those from serial recall, with strong primacy and weak recency (Ward et al., 2010), suggesting they are attempting to recall the entire list in order. Such a strategy is not required when a single item is required for recall, which may explain the weaker primacy under such recall requirements (Tan et al., 2016; Ward & Tan, 2019).

Another reason that manipulation of recall requirements is theoretically constraining concerns an alternate interpretation of how retrieval operates. While the majority of free recall models assume that only one item is retrieved at-a-time (e.g., Howard & Kahana, 2002; Raaijmakers & Shiffrin, 1981), an alternative possibility is that multiple items are simultaneously retrieved—but output sequentially—and retrieval time is longer for larger groups of retrieved items (e.g., Dennis, 2009). If larger groups of items are retrieved when participants initiate at the beginning-of-the-list, slower RTs would be produced, which would resemble the predictions of the separate gradients model. Osth and Farrell investigated this possibility by defining group size as the number of items recalled that were adjacent to each other in the study list, beginning with the first recall (e.g., recalling items 1, 2, 3, 5, and 7 would be a group size of 3). While initiation with the first item was associated with larger group sizes, there was little relationship between group size and first recall latency, arguing against this alternate interpretation. Manipulation of recall requirements provides a more direct test of this possibility that avoids an arbitrary specification of group sizes.

In terms of recency, prior studies have found support for power law functions under similar experimental parameters when only a single item is required at retrieval. Donkin and Nosofsky (2012a) found evidence for power functions in single item recognition where participants were required to study only 12 items. Thus, one possible reason why Osth and Farrell's support for an exponential recency function stands out compared to the literature favoring power functions is that free recall requires many items for output whereas the item recognition task requires retrieval of only a single item.

In the present work, we tested whether recall requirements were responsible for the findings of Osth and Farrell. We conducted an experiment where we presented participants with either six- or 12-item lists (manipulated between subjects)[2] and subsequently prompted them to recall as many items as possible (the *all* condition, which is the standard free recall instruction) or to only recall a single item from the list (the *one* condition) and RTs were recorded from participants' typed responses. The list length of six (LL-6) was within the range of list lengths investigated in previous studies by Ward and colleagues. We additionally

[2] The initial submission of this manuscript contained only 12-item lists. The condition with six-item lists was run at the suggestion of a reviewer where data was collected online due to the impact of COVID-19.

used a 12-item lists (LL-12) as this is closer to the list lengths investigated by Osth and Farrell—longer lists show both weaker primacy and stronger recency effects (Ward et al., 2010).

To investigate whether the recall requirements induced a qualitative change in the nature of primacy and recency mechanisms, we applied a number of LBA models to the data comparing primacy mechanisms (combined vs. separate gradients models) and recency functions (exponential vs. power law functions) for each condition. In addition, we contrasted these models with relatively simple pure-primacy and pure-recency models. All models were implemented in a hierarchical Bayesian framework. If start-of-list reinstatement is selectively adopted as a strategy under standard free recall instructions, then we should only see preference for the separate gradients model in the "all" condition while the "one" condition should show evidence for a combined model. If the slower observed RTs were due to the retrieval of groups of items, then the "one" condition should show relatively constant RTs for each serial position in contrast to the "all" condition. Either finding would be of theoretical significance, as it would reveal that our understanding of primacy is obfuscated by standard free recall instructions to recall many items.

Likewise, if power law recency functions are found when only a single item is required for retrieval, we may find evidence for a power law function in the "one" condition while finding evidence for an exponential function in the "all" condition. This result would be evident in the PFR curves—while the recency effect in the "all" condition should be strongly peaked with an absence of midlist recalls, support for the power law function in the "one" condition would be evident if recalls drop off gradually as lag from the end-of-the-list is increased, showing higher proportions of mid-list recalls. Such a result would accord with the findings from other memory tasks that show evidence for power law functions (Donkin & Nosofsky, 2012b; Rubin & Wenzel, 1996).

Nonetheless, it remains possible that the primacy-to-recency shift observed by Ward and colleagues is not due to qualitative differences, such as changes in recall strategies, in which case we should find evidence for the same models across the recall requirements conditions. In this circumstance, the LBA provides the opportunity to provide some insight into the effects of the manipulation by comparing parameter estimates across the two conditions to evaluate whether the primacy-to-recency shift is due to changes in thresholds, drift rates, the probability of initiating recall with the primacy gradient, or some combination thereof.

Because evidence accumulation models such as the LBA require more data than typical free recall experiments to provide stable estimates of the model parameters, we maximized the number of trials in two ways. First, the majority of the participants completed two 1-h sessions. Second, in the "all" condition participants were given relatively short time periods for recall (up to 25 s) and were allowed to terminate the trial if they were unable to recall any additional words. Optional termination has been used in previous free recall studies (Dougherty et al., 2014; Harbison et al., 2009). A direct comparison between free recall with and without optional termination found no significant differences in the proportions of recalled items between the two procedures—the only observed difference was a longer RT for the last recalled item under optional termination (Hussey et al., 2014). Our two design choices resulted in a total of 180 trials per participant (90 per condition). For comparison, Experiment 1 of the Pennsylvania Electrophysiology of Encoding and Retrieval Study (PEERS: Healey & Kahana, 2014, 2016; Lohnas et al., 2015), which is among one of the larger free recall studies, collected a total of 96 trials across six sessions.

Method

Participants

Participants were 87 undergraduate students from the University of Melbourne, of whom 20 were paid 10$ per session (all in LL-6) for participation, while the remainder received course credit for their participation. Of these participants, 45 and 42 contributed to LL-6 and LL-12, respectively. Four participants (two from each LL) did not complete the second session.

Materials

A word pool of 1409 words with between four and nine letters ($M = 5.68$, $SD = 1.09$) from the N-Watch (Davis, 2005) database. The words were between 25 and 400 counts per million in CELEX word frequency counts ($M = 80.31$, $SD = 69.67$).

Procedure

The LL-12 condition was programmed in Python using the SMILE package (https://github.com/compmem/smile) and administered in a lab. Due to the impact of COVID-19, the LL-6 condition was administered online and programmed using jsPsych and was run at a later time at the request of a reviewer (de Leeuw, 2015).

Each session began with four practice trials followed by 90 experimental trials. Participants were presented with the instructions before both the practice and experimental sessions and were informed that the practice trials would not be scored. Participants were instructed to not to worry about spelling errors and to type the words naturally. For both the practice and experimental trials, each recall requirement condition occupied half of the total trials.

During the study phase, participants were presented with a list of either six or 12 words presented one at a time for 1 s per word followed by a 200-ms interstimulus interval. After the list was presented, participants were presented with a cue for their recall requirements. A text box with the word "ALL" or "ONE" was presented above where the words were presented. Five-hundred milliseconds after the condition prompt appeared, a prompt appeared for participants to begin typing their responses. In the LL-12 condition, this took the form of a string of X's ("XXXXX") which was replaced by the participant's keystrokes, while this was a text box in the LL-6 condition. The characters the participant typed appeared on the screen until they hit the "ENTER" key. In the "one" condition, this resulted in the termination of the trial and the onset of the next trial. In the "all" condition, the prompt reappeared and participants could begin typing the next word. In both conditions, the trial ended either when the participant typed the word "done" and hit the "ENTER" key or until the time limit had expired. The time limit was 15 and 25 s for the "one" and "all" conditions, respectively. Both sessions of the experiment were identical.

Due to experimenter error, some study lists in the LL-12 condition had either a blank presentation or the string "cent200" in place of one of the words. These trials were excluded from the data. This resulted in the exclusion of no more than two trials for each participant ($M = 1.57$, less than 1% of the data). Additionally, in the LL-6 condition there was an error where for the first 12 participants the final study list was not tested. This was corrected such that all remaining participants were tested on all 90 lists.

Data processing

All erroneous responses were spell-checked. Spell-checking was semi-automated using a computer program. First, the string similarity between erroneous responses and each word in the set was calculated using the Levenshtein–Damerau distance divided by the longer of the two words. Responses were only considered for spell-checking if the similarity between the response and another word was at least .50. Subsequently, the response and the pool of matching words were presented to the user (the first or second author). The error was only spell-checked if the error was not a word in its own right. For instance, if the participant typed "miles" and one of the matching words was "smiles", the error was not replaced.

Computational modeling

LBA models of primacy and recency

A total of seven LBA models were applied to the data: four models were constructed by factorially crossing the recency functions (exponential and power law) and primacy mechanisms (combined vs. separate gradients), in addition to the pure-primacy model and two pure-recency models (power vs. exponential). All seven models shared common decision-related parameters, including starting point variability A, the distance between the height of the starting point distribution and the response threshold B, and the time for non-decision processes t_0, which is added to the predicted RT distribution.

Recency functions

Recency functions were implemented as drift rate functions. Each recency function consists of a scale parameter α and decay parameter β. In the combined model, α varies across serial positions to implement primacy (described in the next section). Instead of estimating α, we estimate a parameter a—α is proportional to a when primacy combines with the recency function, otherwise α is identical to a. Following Osth and Farrell (2019), we did not include an asymptote parameter in the functions. The drift rate v for serial position i in the exponential function is determined as follows:

$$v_i = \alpha_i e^{-\beta t} \tag{1}$$

where t is the study-test lag (measured as $L - i$, with L being the number of items on the list).

The power law function is written as:

$$v_i = \alpha_i t^{-\beta} \tag{2}$$

As mentioned previously, several free recall models capture recency using a context representation that gradually changes during the study list, matching the more recent items on the list when used as a cue (Howard & Kahana, 2002; Mensink & Raaijmakers, 1988). If context changes more rapidly, it produces more "peaked" recency functions that better match the later items on the list. Thus, β can be interpreted as the rate of contextual change, while a may correspond to the strength of the item-context associations being formed. While both parameters appear to be linked to encoding, there are alternative interpretations that may allow such parameters to vary at retrieval. If there are multiple context representations that change at different rates (Howard et al., 2015; Pashler et al., 2009), it may be possible for participants to select one of several context representations at retrieval. We return to this possibility in the General Discussion.

Primacy mechanisms

The combined model assumes that the beginning of list items receive a strength boost, resulting in a single non-monotonic drift rate function. We implemented this by applying an exponentially decreasing function on the α parameters for each serial position:

$$\alpha_i = a(r_s \exp[-r_d(i-1)] + 1) \tag{3}$$

where r_s and r_d are the primacy scale and decay parameters, and a is the base recency scale parameter. If r_s is zero, α_i reverts to a. If r_d is sufficiently large, it produces a relatively steep primacy gradient such that α_i will resemble a for the later list items.

In the separate gradients mixture model, in contrast, participants employ different drift rate functions on different trials. Specifically, the recency function is either Eqs. 1 or 2 where $\alpha_i = a$ and the primacy function is:

$$v_i = r_s \exp[r_d(i-1)] \tag{4}$$

This model additionally uses a mixing parameter p. On a proportion of trials p, participants employ the primacy gradient specified in Eq. 4 as their drift rate function which reflects start-of-list reinstatement. On proportion $1 - p$ trials, in contrast, participants rely on an end-of-list context, employing a recency drift rate function specified by Eqs. 1 or 2, depending on whether the model specifies an exponential or power function. For the pure-primacy model, only the primacy function of Eq. 4 is employed. For the pure-recency models, $\alpha_i = a$ and only Eqs. 1 or 2 is employed.

Interpretation of r_s and r_d depend on the model. In the combined model, r_s can be interpreted as the boost to the early list items, which can come from sources such as decreasing attention, temporal distinctiveness, or extra time in a rehearsal buffer, while r_d might govern the extent to which these benefits extend to later items. In the separate gradients model, r_s can be interpreted as the strength of the reinstated start-of-list context while r_d might represent the extent to which the start-of-list is context is associated to later items.

The model fit

Six participants were excluded for showing high error rates in their first responses (LL-6: 28.3%, 16.6%, 12.7%, LL-12: 22.3%, 12.9%, 14.6%). A high proportion of these errors were prior-list intrusions in five of the participants (LL-6: 3.9%, 66.6%, 17.4%, LL-12: 42.5%, 21.7%, 15.4%). Trials where participants failed to recall a single word were excluded – this amounted to less than 1% of the data. We also excluded trials where recall was initiated with a response other than one of the list words (4.6% of trials).

We used the RT of the first keypress relative to the onset of the condition cue ("one" vs. "all") of each word response as our latency measure. This resembles the usage of the onset of each word when vocal responses are recorded (Murdock & Okada, 1970).

Each model was applied to the "all" and "one" conditions separately. Seven parameters are estimated for each combined model (t_0, A, B, a, β, r_s, and r_d). The pure-primacy model omits a and β while the pure-recency models omit the r_s and r_d parameters, resulting in five parameters for each model. Eight parameters were estimated for each separate gradients model (which use the same parameters as the combined model, but also uses the mixing parameter p).

To avoid distortions associated with fitting group-level data, the models were fit to data using hierarchical Bayesian techniques (Rouder & Lu, 2005; Shiffrin et al., 2008). Similar to maximum likelihood estimation (MLE), each individual response (RT and serial position) is fit and separate parameters are allotted for each participant. However, hierarchical Bayesian methods depart from MLE in two important ways. First, a separate *group-level* distribution is estimated—in addition to estimating the likelihood of a participant's data under their own parameters, the likelihood of the individual participant's parameters is estimated according to the group-level distribution. This "pools" across individuals; the estimation of a participant's parameters is affected by the other participants in the dataset. Second, rather than using point-estimates for each parameter, Bayesian methods allow for quantifying uncertainty in the parameters as posterior distributions.

Parameters were estimated using differential-evolution Markov chain Monte Carlo (DE-MCMC) techniques (Turner et al., 2013), which are robust to parameter correlations. Minimally-informative prior distributions were employed on all group-level distributions. Details can be found in Supplementary Materials C.

Model selection

The models vary in their degree of complexity, with the pure-primacy and recency models employing the fewest parameters and the separate gradients models employing the most parameters. For this reason, we selected between models using the widely applicable information criterion (WAIC: Watanabe, 2010). Similar to other information criteria such as AIC and BIC, WAIC calculates a balance between goodness-of-fit and model complexity, but differs in that it is an asymptotic approximation of leave-one-out cross validation. More complex models receive harsher penalties, thus a complex model has to justify its additional complexity with a greater increase in its goodness-of-fit. Conventionally, WAIC differences between models of 10 or more are considered “large”. In addition, we also report WAIC weights, in which a ratio of transformed WAIC values divided by their sum (Wagenmakers & Farrell, 2004). Under certain assumptions, WAIC weights can be interpreted as the probability that a given model is the data-generating model for the dataset among the set of models under consideration.

Model selection results are depicted in Table 1 using WAIC difference scores, where the winning model equals zero and all positive values are WAIC penalties relative to the winning model. Both the “all” and “one” conditions in both list length conditions reveal the same winner: an exponential recency gradient where primacy is instantiated using the separate gradients mixture model. The advantages for this model are quite large in each case, with advantages of over 50 points over the next best model in each comparison. WAIC weights for the winning models are decisive (1.0), and indicate confidence that they are providing substantially better account of the data than the other models.

While the results of the “all” condition in the LL-12 dataset replicate the Osth and Farrell (2019) model selection results, the results of the “one” condition are novel and suggest the same decision dynamics underlie retrieval of a single item as multiple items—we found no evidence for qualitative changes in recency functions and primacy mechanisms across the two recall requirements conditions. In addition, this provides the first demonstration of an advantage for exponential recency gradients and the separate gradients mixture account of primacy with lists as short as six items in length. In the next section, we will discuss the data and model fits to analyze the success of this model.

Serial position curves

Figure 2 shows group-averaged serial position curves (SPCs: row A) and probability of first recall (PFR) curves (row B) for all conditions. The SPCs depict the probability of each serial position being recalled over the entire sequence of recalls for the “all” condition. The comparison to the PFR curves in row B demonstrates that the SPCs show several of the same trends—primacy is dominant for the LL-6 condition, whereas a small primacy and large recency effect can be found in the LL-12 condition. Additionally, the SPCs reveal that the short recall times and the option to terminate responding did not dissuade participants from recalling a reasonable number of items from each list. The fact that both primacy and recency effects are present in the PFR curves in all conditions is likely the reason why the pure-primacy and pure-recency models performed poorly in the model selection.

The data reproduce the primacy-to-recency shift as recall requirements are lightened (Tan et al., 2016; Ward & Tan, 2019). Both list lengths show a reduction in primacy and increase in recency in the “one” condition relative to the “all” condition. However, this pattern was much more pronounced in the LL-6 condition, which demonstrated a large decrease in PFR for the first item (M_{all} = .56, SEM_{all} = .046, M_{one} = .28, SEM_{one} = .034) as well as a large increase in PFR for the final item (M_{all} = .12, SEM_{all} = .028, M_{one} = .35, SEM_{one} = .034). The LL-12 conditions, in contrast, showed a comparatively small decrease in PFR for the first item (M_{all} = .087, SEM_{all} = .015, M_{one} = .058, SEM_{one} = .01) along with a moderate increase in PFR for the final item (M_{all} = .288, SEM_{all} = .039, M_{one} = .419, SEM_{one} = .039).

Table 1 WAIC difference scores for each model, separated by list length and recall requirements conditions (“one” vs. “all”). The winning model is depicted in bold and the WAIC weight is depicted in parentheses

			All			One		
			Recency			Recency		
LL	*N*	Primacy	None	Exp	Power	None	Exp	Power
6	5	Pure-Primacy	852 (0)	-	-	2065 (0)	-	-
	5	Pure-Recency	-	4622 (0)	4641 (0)	-	1824 (0)	1846 (0)
	7	Combined	-	230 (0)	469 (0)	-	331 (0)	670 (0)
	8	Separate	-	**0 (1.0)**	54 (0)	-	**0 (1.0)**	106 (0)
12	5	Pure-Primacy	3510 (0)	-	-	5542 (0)	-	-
	5	Pure-Recency	-	674 (0)	944 (0)	-	492 (0)	703 (0)
	7	Combined	-	357 (0)	751 (0)	-	371 (0)	673 (0)
	8	Separate	-	**0 (1.0)**	212 (0)	-	**0 (1.0)**	89 (0)

Notes: LL = list length, *N* = number of individual participant parameters in the model

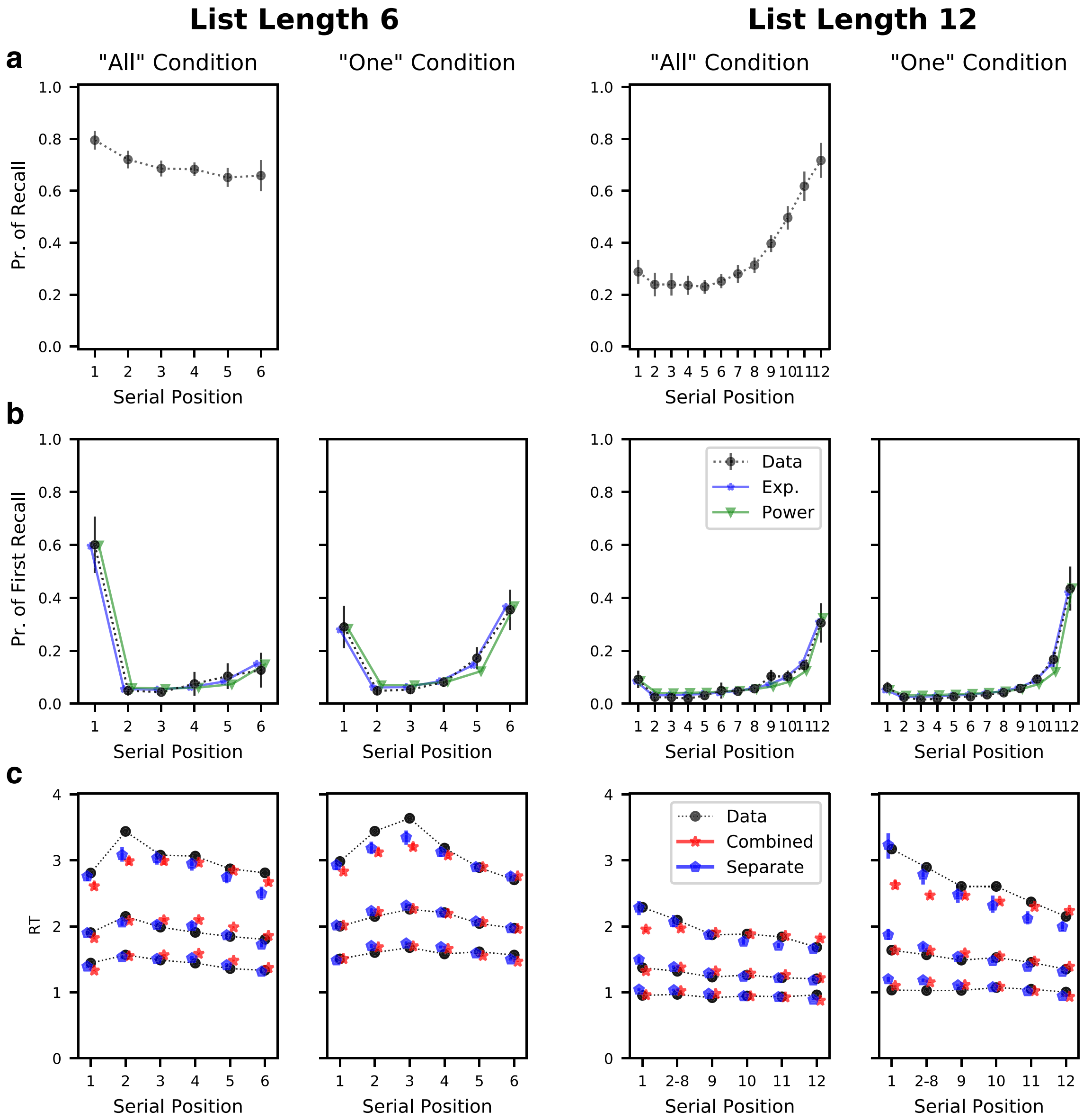

Fig. 2 Group-averaged serial position curves (row A), probability of first recall curves (row B), and RT distributions (row C) for all conditions, along with posterior predictives from a selection of relevant LBA models (see text for details). The RT distributions summarized by the .1, .5, and .9 quantiles for the data (which were smoothed by a hierarchical ex-Gaussian model) along with the winning combined and separate gradients models of primacy. *Error bars* in row A and B are 95% within-subjects confidence intervals, while the error bars in row C are 95% highest density intervals (HDIs) from the LBA models

Osth and Farrell (2019) noted that both the best combined and separate gradients models provided very good descriptions of the PFR curves. While PFR curves do not discriminate between these models, they do discriminate between the recency functions. For this reason, Fig. 2B depicts the winning exponential and power law models for the "all" and "one" conditions. Despite the differences between these conditions, the exponential recency gradient appears to provide a better account of the data than the power law gradient. The differences between the two models is subtle, but the exponential gradient is better able to capture the recalls in the last four serial positions than the power law gradient, as the power law gradient predicts a more gradual fall-off in recall probability with time than is shown in the data. We were quite surprised to find that the advantage for the exponential function even applies to the LL-6 condition. While both functions appear very similar in the LL-6 condition, the exponential function appears to provide a better fit to the 5th serial position across both the "one" and "all" conditions. Interestingly, the strength of the exponential recency gradient's account is especially pronounced in the "one" condition where it provides an excellent account of the data. Thus, these results provide little support for the idea that retrieval of a single item qualitatively changes the nature of the recency gradient.

There is one point of misfit of both models in the "all" requirement in the LL-12 dataset—the 9th item was recalled more frequently than either of the two models predict. Supplementary Materials A shows the fits to the individual participants from the winning model. Similar to previous work examining individual differences (e.g., Healey & Kahana, 2014), there was considerable variability in PFR curves, with participants demonstrating either mostly primacy, mostly recency, a combination of primacy and recency, along with a small subset of participants who initiated recall with an item that was between one and five positions before the final item, a pattern that was more pronounced in the "all" condition. Healey and Kahana (2014) found a similar subgroup of participants in their dataset and attributed it strategies such as forming "groups" of items and imitating recall with the first member of the group (e.g., Farrell, 2012). There is currently no mechanism for temporal grouping in our models, and while such a mechanism is possible to include, it goes considerably beyond the scope of the present work.

RT distributions

While the PFR curves are discriminating among the recency functions, the extent to which the RT distributions vary by serial position is discriminating among the primacy mechanisms. RT distributions were summarized using the .1, .5, and .9 quantiles, which are the 10th, 50th, and 90th percentiles. A difficulty in evaluating the fit to RT distributions is that there are very sparse numbers of observations from some serial positions, with some participants showing few or no recalls from some positions, preventing estimation of the .1 and .9 quantiles.

To remedy these problems, following Osth and Farrell (2019), in the LL-12 dataset we combined all of the mid-list positions (serial positions two through eight) into a single bin due to their sparse recall counts. In addition, we parametrically smoothed each participant's RTs using a hierarchical ex-Gaussian model (details of this procedure are described in Supplementary Materials C). The ex-Gaussian distribution has been demonstrated to give an excellent account of free recall RT distributions (Wixted & Rohrer, 1993, 1994; Unsworth & Engle, 2007). This procedure was performed solely to summarize the data and to evaluate the fit of the models.

Group-averaged RT quantiles from the smoothed data along with the predictions of the combined and separate gradients models of primacy (each of which use an exponential recency gradient) can be seen in Fig. 2C. Only participants with a minimum of two recalls in each bin were included in the plot. In each condition, the .1 quantile is fairly constant across serial positions. However, both the median and especially the .9 quantile show greater changes across serial positions. A peculiar finding is that RTs are shorter in the longer list length (LL-12). We hesitate to make any strong claims about this difference as the LL-6 dataset was collected online using different experiment software, which may have resulted in slower RTs.

For the LL-12 dataset, both recall requirements conditions show the slowest RTs for the first serial position. The key difference between the conditions, surprisingly, was that RTs were slowest in the "one" condition. The posterior predictives from each of the models largely reflect those shown in the initial predictions in Fig. 1. Specifically, the combined model shows relatively constant latencies for each serial position while the separate gradients model is better able to capture the RT differences across serial positions and captures the slower RTs for the earlier serial positions.

For the LL-6 dataset, the RT differences across serial positions and recall requirements are more subtle. In particular, the first item has similar RT to the final item and the "one" and "all" conditions also have similar RTs. In this dataset, the separate gradients model yields its advantage because it is better able to capture the RT variability across serial positions, particularly the relatively slow RT of the second item. Analysis of the drift rate functions constructed from the parameter estimates in the next section provide some insight as to why the shorter list provided a different pattern of RTs across serial positions.

Fits to individual participants' RT distribution with the winning model can be found in Supplementary Materials A, where it is revealed that the model provides a convincing account of the RT differences across individuals in all conditions.

Parameter estimates

The results of the modeling suggest that changes in recall requirements do not qualitatively change the primacy mechanism or the shape of the recency function. The question then is: what is affected by the recall requirements manipulation? Specifically, what causes the greater emphasis on recency when only a single item is required for output? To address this question, the top row of Fig. 3 compares the primacy and recency functions constructed from the group mean parameters of the winning model across the two recall requirements conditions, while rows 2–4 depict the posterior distributions and the 95% highest density intervals (HDIs) of the group mean parameters.

Despite the large reductions in primacy in the "all" condition of the LL-6 dataset, the primacy function appears to be unaffected by the manipulation. Instead, the reduction in primacy is likely to have arisen from the large reduction

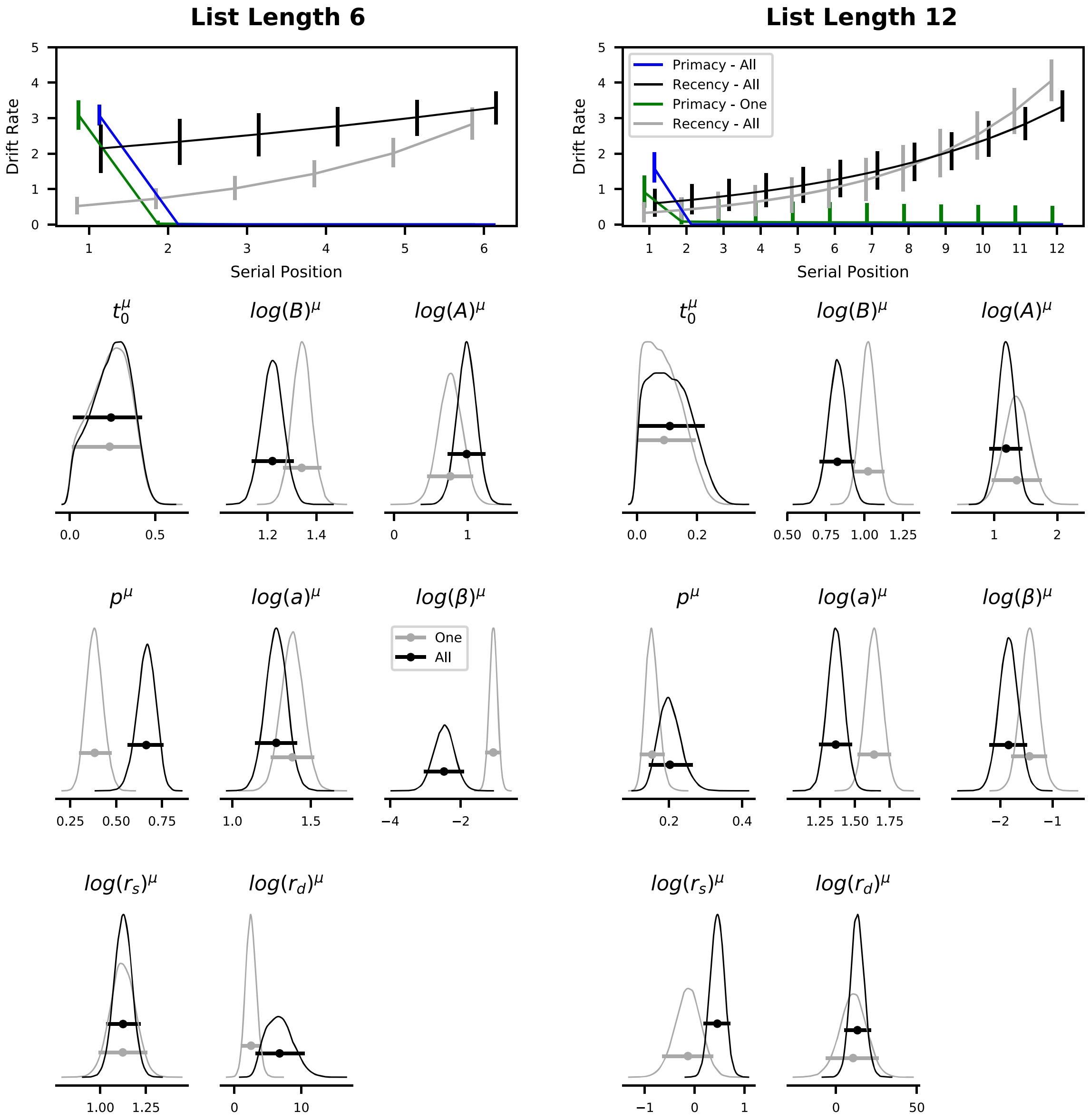

Fig. 3 The *top row* shows the primacy and recency functions constructed from the group mean parameters for each condition. *Error bars* depict the 95% highest density intervals (HDIs). Rows 2–4 show the posterior distributions and 95% HDIs of the group mean parameters across the “one” and “all” conditions. Some parameters on a (0, ∞) scale are log-transformed - see Supplementary Materials C for details

in the mixture parameter p^μ, suggesting that participants are less likely to reinstate the start-of-the-list in the “one” condition. In addition, the recency function appears more “peaked” in the “one” condition, which is reflected in a higher value of the recency decay rate parameter $log(\beta)^\mu$. The drift rate functions also show similar maximum drift rates for the primacy and recency functions, which reveals why the RTs in Fig. 2 did not vary considerably by serial position. In short lists, primacy items are more recent than in longer lists, which may be why they are of comparable strength.

The LL-12 dataset similarly shows a more peaked recency function in the “one” condition as well as a trend toward higher values of $log(\beta)^\mu$. This is accompanied by a higher maximum drift rate, reflected in a higher value of the recency scale parameter $log(a)^\mu$, suggesting that the stronger recency effect comes from higher drift rates in the recency function. The reduction in primacy appears to come from trends in two parameters—a weaker maximum drift rate in the primacy function (reflected in a lower value of $log(r_s)^\mu$) and a lower value of the mixture parameter p^μ. Both LL6 and LL-12 datasets show trends for higher values of the retrieval threshold ($log(B)^{\mu}$) in the "one" condition as well.

Nonetheless, the uncertainty in the group mean parameter estimates was somewhat large. One possible reason for this is that all parameters were allowed to vary across the two recall requirements conditions. We attempted to find a simpler account of the two conditions by fitting both conditions simultaneously and constraining some parameters to take the same value across the "one" and "all" conditions. To keep to a manageable number of models, t_0 and A did not vary across conditions. We then pursued three classes of differences, and combinations thereof: 1.) a threshold model, where the B parameter varied across the two conditions, 2.) a drift rate model, where the scale parameters of the recency (a) and primacy (r_s) gradients varied, and 3.) a cuing model, where the mixture parameter of the primacy gradient p varied. These explorations were restricted to the winning model from the previous model selection, namely the separate gradients mixture model with an exponential recency gradient.

Each of the models were compared along with the full model, where all parameters were allowed to vary across conditions, in Table 2. Results are summarized using WAIC difference scores and WAIC weights. One can see that the full model wins decisively, suggesting that all parameters were affected by the manipulation of recall requirements. Why were all parameters affected? Supplementary Materials B shows some of the fits of the alternative models. As it turns out, the cuing only model was able to reasonably capture the group-averaged PFR curves in all conditions, but was not able to capture the RT differences across conditions. The threshold + cuing model, in contrast, provided a good account of the qualitative trends in the group-averaged data was better able to capture the RT differences by allowing a higher value of B in the "one" condition. However, additional mechanisms were required to capture some of the variation across individual participants. For instance, in the LL-12 dataset, there are some participants who show pure recency patterns that are more sharply peaked in one condition than another. This transition was only able to be captured by higher values of a in that condition and additionally benefit by variation of β in the full model.

General discussion

In this work, we explored the dynamics of decision-making in free recall initiation when recall requirements were manipulated, such that either the entire list or only a single item was required for retrieval. We explored this paradigm because previous work has demonstrated that more lenient recall requirements result in a shift from primacy-to-recency (Tan et al., 2016; Ward & Tan, 2019). If less items are required for output, it is possible that participants are less likely to reinstate the beginning-of-the-list, as cuing from the beginning of the list will be geared towards recalling a larger number of items. We explored this issue using two different list lengths – six items (which was used by Ward and colleagues) along with a longer list of 12 items, which is closer to the list lengths employed in the investigation of Osth and Farrell.

We used the LBA model of decision-making to implement various models of free recall initiation and jointly fit the serial position and RT distributions of recalls. Reinstatement of the start-of-the-list in the LBA can be modeled by assuming a mixture of primacy and recency gradients, such that participants rely on different drift rate gradients on different trials (the separate gradients mixture model). This qualitatively represents models where participants form associations to a start cue and reinstate it to provide a different cue to retrieve the list items (e.g., Farrell, 2012; Metcalfe & Murdock, 1981; Morton & Polyn, 2016). This stands in contrast to models where primacy and recency items jointly compete to be retrieved with differences in strength, or drift rate, being the only

Table 2 WAIC difference scores for each model. The winning model is depicted in bold and the WAIC weight is depicted in parentheses

Model	N	LL-6	LL-12
Threshold	9	1745 (0)	829 (0)
Drift	10	1172 (0)	596 (0)
Cuing	9	384 (0)	789 (0)
Threshold + Drift	11	940 (0)	221 (0)
Threshold + Cuing	10	268 (0)	492 (0)
Drift + Cuing	11	221 (0)	390 (0)
Threshold + Drift + Cuing	12	92 (0)	196 (0)
All	16	**0 (1.0)**	**0 (1.0)**

Notes: N = number of individual participant parameters per model, LL = list length factor that differentiates them (the combined model, e.g., Atkinson & Shiffrin, 1968; Brown et al. 2007; Raaijmakers & Shiffrin, 1981).

Our model selection results strongly suggested the same mechanism, namely start-of-list reinstatement, underlies retrieval of a single item and the entire list in both list length conditions. This is likely due to the fact that the combined model of primacy makes rather constrained predictions about RTs, such that similar RTs are predicted for each item, even after the model has been fit to data. The separate gradients model, in contrast, is more flexible in its predictions about RTs and in particular can predict slower RTs for primacy items if the primacy gradient is weaker than the recency gradient. While the LL-6 condition showed similar RTs for each serial position, the separate gradients model was still better able to capture the RT variability across serial positions, particularly the slow RT for the second item. For the LL-12 condition, in contrast, the first item was considerably slower than subsequent items in both recall requirements conditions, and these patterns were best described by the separate gradients model, which attributed this pattern to a mixture of a fast recency race and a relatively slow primacy race.

An alternative explanation for the slower RTs for primacy items is that multiple items are retrieved simultaneously and output sequentially, with RT being proportional to the size of the group (Dennis, 2009). If participants are more likely to retrieve groups of items when they initiate at the beginning, this would result in slower RTs for primacy items. However, this alternative explanation is undermined by the finding that slower RTs for the primacy items in the LL-12 dataset were found even in the “one” condition, where multiple items were not required for retrieval. While it is possible that participants covertly recalled multiple items and only output the first item, it is unclear why participants would otherwise show stronger recency and slower RTs in the “one” condition.

Why would participants reinstate the beginning-of-the-list when only a single item is required for recall? One possibility is that start-of-list reinstatement may occur automatically (but randomly) as a consequence of a reminder of the event, namely the study list. If this is the case, then reinstatement may occur regardless of whether a single item or multiple items are required for retrieval. This does not necessarily imply that participants have no control over start-of-list reinstatement, as it is likely required for initiating recall at the beginning of the list, which is required for serial recall. Start-of-list reinstatement is common in models of serial recall, where items are associated to their ordinal positions and the position of the first item is retrieved to initiate recall (e.g., Anderson & Matessa, 1997; Brown et al., 2000; Farrell, 2012; Henson, 1998). Evidence for such positional representations comes from the finding that when participants make intrusions from prior lists, they tend to be retrieved in the same output position as they were recalled in the prior list (Henson, 1998; Osth & Dennis, 2015; Fischer-Baum & McCloskey, 2015).

In addition to varying the primacy mechanisms, we additionally explored whether the recency gradient in each condition was an exponential or power function. Osth and Farrell (2019) found strong evidence for exponential recency gradients, a result which is seemingly at odds with a large literature favoring power functions of forgetting (e.g., Rubin & Wenzel, 1996; Wixted & Ebbesen, 1991). One investigation has even supported power functions under similar experimental parameters (lists of 12 items) required retrieval of a single item (e.g., Donkin & Nosofsky, 2012b), namely single item recognition. In our investigation, the exponential recency function yielded a superior account of the PFR curves even in the “one” condition in both list lengths because the exponential function was better able to account for the decreasing recall probabilities for items prior to the final item. Thus, our investigation instead suggests that the divergence from the prior literature may be due to differences between the memory tasks, and not due to differences in the required number of to-be-retrieved items.

Given the lack of qualitative differences between conditions, what is responsible for the primacy-to-recency shift as recall requirements are relaxed? Both our parameter comparison and model selection found that several LBA parameters were affected by the recall requirements manipulation, including a trend toward higher decision thresholds as well as a reduction in cuing with the start-of-the-list (as reflected by the p parameter) which was especially pronounced in the LL-6 dataset. In the LL-12 dataset, the recency function became more peaked as recall requirements were lightened. But how can existing memory models produce stronger drift rates for recency items in response to an instructional manipulation? In several models, recency is the result of a drifting context representation that better matches the later items at retrieval (e.g., Davelaar et al., 2005; Howard & Kahana, 2002; Mensink & Raaijmakers, 1988; Osth et al.. 2018). However, in some models there are multiple context representations that change at different rates, corresponding to different timescales (e.g., Howard et al., 2015; Pashler et al., 2009)—participants may be able to selectively use a context representation corresponding to a shorter timescale that is represented more strongly. Multiscale context models are also capable of producing power law functions, as an average of a number of exponential functions with different decay rates can approximate a power law function (Anderson & Tweeney, 1997). Osth and Farrell conjectured that the evidence for exponential recency functions in free recall may be due to participants selectively using one context representation for retrieval. Evidence for the ability to retrieve timescales of interest comes from autobiographical recall, where participants are able to retrieve items from particular time periods instructed by the experimenter (Moreton & Ward, 2010).

Supplementary Information The online version contains supplementary material available at (10.3758/s13421-020-01117-2).

References

Anderson, J., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*, 728–748.

Anderson, R. B., & Tweeney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*, 724–730.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, *2*, 89–195.

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*(1), 25–35.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*, 127–181.

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

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *36*(2), 484–499.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review*, *112*(1), 3–42.

Davis, C. J. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*(1), 65–70.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*, 1–12.

Dennis, S. (2009). Can a chaining model account for serial recall?. In Carlson, L., Hülscher, C., & Shipley, T. (Eds.) *Proceedings of the XXXI Annual Conference of the Cognitive Science Society*, (pp. 2813–2818).

Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science*, *23*(6), 625–634.

Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: an investigation using response time distribution models. *Psychonomic Bulletin & Review*, *19*, 363–394.

Dougherty, M. R., Harbison, J. I., & Davelaar, E. J. (2014). Optional stopping and the termination of memory retrieval. *Current Directions in Psychological Science*, *23*, 332–337.

Evans, N. J., & Wagenmakers, E.-J. (2020). Evidence accumulation models: current limitations and future directions. *The Quantitative Methods for Psychology*, *16*, 73–90.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271.

Fischer-Baum, S., & McCloskey, M. (2015). Representation of item position in immediate serial recall: Evidence from intrusion errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1426–1446.

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language, 110*.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition*, 397–402.

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies?. *Journal of Experimental Psychology: General*, *143*, 575–596.

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

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, *26*, 699–720.

Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology*, *36*, 73–137.

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), 923–941.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 268–299.

Howard, M. W. (2004). Scaling behavior in the temporal context model. *Journal of Mathematical Psychology*, *48*(4), 230–238.

Howard, M. W., Shankar, K. H., Aue, W. R., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*(1), 24–53.

Hussey, E. K., Dougherty, M. R., Harbison, J. I., & Davelaar, E. J. (2014). Retrieval dynamics in self-terminated memory search. *The Quarterly Journal of Experimental Psychology*, *67*, 394–416.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*, *24*, 103–109.

Laming, D. (1999). Testing the idea of distinct storage mechanisms in memory. *International Journal of Psychology*, *34*, 419–426.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*(1), 155–189.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337–363.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.

Mensink, G. J., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*(4), 434–455.

Metcalfe, J., & Murdock, B. B. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior*, *20*, 161–189.

Moreton, B. J., & Ward, G. (2010). Time scale similarity and long-term memory for autobiographical events. *Psychonomic Bulletin & Review*, *17*, 510–515.

Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. *Journal of Memory and Language*, *86*, 119–140.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, *86*(2), 263–267.

Osth, A. F., & Dennis, S. (2015). Prior-list intrusions in serial recall are positional. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1893–1901.

Osth, A. F., Bora, B., Dennis, S., & Heathcote, A. (2017). Diffusion versus linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory and Language*, *96*, 36–61.

Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology*, *104*, 106–142.

Osth, A. F., & Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological Review*, *126*, 578–609.

Pashler, H., Cepeda, N., Lindsey, R. V., Vul, E., & Mozer, M. C. (2009). Predicting the optimal spacing of study: a multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta (Eds.), *Advances in Neural Information Processing Systems* 22 (pp. 1321–1329). Curran Associates, Inc.

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

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*, 574–590.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of Associative Memory. *Psychological Review*, *88*(2), 93–134.

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

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential-sampling models for two choice reaction time. *Psychological Review*, *111*, 333–367.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

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

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.

Serruya, M. D., Sederberg, P. B., & Kahana, M. J. (2014). Power shifts track serial position and modulate encoding in human episodic memory. *Cerebral Cortex*, *24*, 403–413.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: a foundational primer. *Journal of Mathematical Psychology*, *44*, 408–463.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of the zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*, 1–34.

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), 1589–1625.

Tan, L., Ward, G., Paulaskaite, L., & Markou, M. (2016). Beginning at the beginning: Recall order and the number of words to be recalled. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1282–1292.

Tillman, G., Van Zandt, T., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review*, *27*, 911–936.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384.

Unsworth, N., & Engle, R. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550–592.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1207–1241.

Ward, G., & Tan, L. (2019). Control processes in short-term storage: Retrieval strategies in immediate recall depend upon the number of words to be recalled. *Memory & Cognition*.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409–415.

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1024–1039.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*(1), 89–106.