Feature and order manipulations in a free recall task affect memory for current and future lists

Jeremy R. Manning[1,*], Emily C. Whitaker[1], Paxton C. Fitzpatrick[1], Madeline R. Lee[1], Allison M. Frantz[1], Bryan J. Bollinger[1], Darya Romanova[1], Campbell E. Field[1], and Andrew C. Heusser[1,2]

[1]Dartmouth College

[2]Akili Interactive Labs

[*]Corresponding author: jeremy.r.manning@dartmouth.edu

Abstract

We perceive, interpret, and remember our ongoing experiences through the lens of our prior experiences. Inferring that we are in one type of situation versus another can lead us to interpret the same physical experience differently. In turn, this can affect how we focus our attention, form expectations about what will happen next, remember what is happening now, draw on our prior related experiences, and so on. To study these phenomena, we asked participants to perform simple word list-learning tasks. Across different experimental conditions, we held the set of to-be-learned words constant, but we manipulated how incidental visual features changed across words and lists, along with the orders in which the words were studied. We found that these manipulations affected not only how the participants recalled the manipulated lists, but also how they recalled later (randomly ordered) lists. Our work shows how structure in our ongoing experiences can influence how we remember both our current experiences and unrelated subsequent experiences.

Keywords: episodic memory, free recall, incidental features, implicit priming, temporal order

Introduction

Experience is subjective: different people who encounter identical physical experiences can take away very different meanings and memories. One reason for this is that our moment-by-moment subjective experiences are shaped in part by the idiosyncratic prior experiences, memories, goals, thoughts, expectations, and emotions that we bring with us into the present moment. These factors collectively define a *context* for our experiences (Manning, 2020).

The contexts we encounter help us to construct *situation models* (Manning et al., 2015; Radvansky and Copeland, 2006; Ranganath and Ritchey, 2012; Zwaan et al., 1995; Zwaan and Radvansky, 1998) or *schemas* (Baldassano et al., 2018; Masís-Obando et al., 2022; Tse et al., 2007) that describe how experiences are likely to unfold based on our prior experiences with similar contextual cues. For example, when we enter a sit-down restaurant, we might expect to be seated at a table, given a menu, and served food. Priming someone to expect a particular situation or context can also influence how they resolve potential ambiguities in their ongoing experiences, including in ambiguous movies and narratives (Rissman et al., 2003; Yeshurun et al., 2017).

Our understanding of how we form situation models and schemas, and how they interact with our subjective experiences and memories, is constrained in part by substantial differences in how we study these processes. Situation models and schemas are most often studied using "naturalistic" stimuli such as narratives and movies (Nastase et al., 2020; Zwaan et al., 1995; Zwaan and Radvansky, 1998). In contrast, our understanding of how we organize our memories has been most widely informed by more traditional paradigms like free recall of random word lists (Kahana, 2012, 2020). In free recall paradigms, participants study lists of items and are instructed to recall the items in any order they choose. The orders in which words come to mind can provide insights into how participants have organized their memories of the studied words. Because random word lists are unstructured by design, it is not clear if or how non-trivial situation models might apply to these stimuli. As we unpack below, this provides an important motivation for our current study, which uses free recall of *structured* lists to help bridge the gap between these two lines of research.

Like remembering real-world experiences, remembering words on a studied list requires distinguishing the current list from the rest of one's experience. To model this fundamental memory capability, cognitive scientists have posited a special context representation that is associated with each list. According to early theories (e.g., Anderson and Bower, 1972; Estes, 1955) context representations are composed of many features which fluctuate from moment to moment, slowly drifting through a multidimensional feature space. During recall, this representation forms part of the retrieval cue, enabling us to distinguish list items from non-list items. Understanding the role of context in memory processes is particularly important in self-cued memory tasks, such as free recall, where the retrieval cue is "context" itself (Howard and Kahana, 2002a). Conceptually, the same general processes might be said to describe how real-world contexts evolve during natural experiences. However, this is still an open area of study (Manning, 2020, 2021).

Over the past half-century, context-based models have had impressive success at explaining many stereotyped behaviors observed during free recall and other list-learning tasks (Estes, 1955; Glenberg et al., 1983; Howard and Kahana, 2002a; Kimball et al., 2007; Polyn and Kahana, 2008; Polyn et al., 2009; Raaijmakers and Shiffrin, 1980; Sederberg et al., 2008; Shankar and Howard, 2012; Sirotin et al., 2005). These phenomena include the well-known recency and primacy effects (superior recall of items from the end and, to a lesser extent, from the beginning of the studied list), as well as semantic and temporal clustering effects (Howard and Kahana, 2002b; Kahana et al., 2008). The contiguity effect is an example of temporal clustering, which is perhaps the dominant form of organization in free recall. This effect can be seen in people's tendencies to successively recall items that occupied neighboring positions in the studied list (Kahana, 1996). There are also striking effects of semantic clustering (Bousfield, 1953; Bousfield et al., 1954; Jenkins and Russell, 1952; Manning and Kahana, 2012; Romney et al., 1993), whereby the recall of a given item is more likely to be followed by recall of a similar or related item than a dissimilar or unrelated one. In general, people organize memories for words along a wide variety of stimulus dimensions. According to models like the *Context Maintenance and Retrieval* model (Polyn et al., 2009), the stimulus features associated with each word (e.g., the word's meaning, size of the object the word represents, letters that make up the word, font size, font color, location on the screen, etc.) are incorporated into the participant's mental context representation (Manning, 2020; Manning et al., 2015, 2011, 2012; Smith and Vela, 2001). During a memory test, any of these features may serve as a memory cue, which in turn leads the participant to successively recall words that share stimulus features.

A key mystery is whether (and how) the sorts of situation models and schemas that people use to organize their memories of real-world experiences might map onto the clustering effects that reflect how people organize their memories for word lists. On one hand, both situation models and clustering effects reflect statistical regularities in ongoing experiences. Our memory systems exploit these regularities when generating inferences about the unobserved past and yet-to-be-experienced future (Bower et al., 1979; Momennejad et al., 2017; Ranganath and Ritchey, 2012; Schapiro and Turk-Browne, 2015; Xu et al., 2023). On the other hand, the rich structures of real-world experiences and other naturalistic stimuli that enable people to form deep and meaningful situation models and schemas have no obvious analogs in simple word lists. Often, lists in free recall studies are explicitly *designed* to be devoid of exploitable temporal structure, for example by sorting the words in a random order (Kahana, 2012).

We designed an experimental paradigm to explore how people organize their memories for simple stimuli (word lists) whose temporal properties change across different "situations," analogous to how the content of real-world experiences changes across different real-world situations. We asked participants to study and freely recall a series of word lists (Fig. 1). In the different conditions in our experiment, we varied the lists' appearances and presentation orders in different ways. The studied items (words) were designed to vary along three general dimensions: semantic (word *category* and physical *size* of the referent), lexicographic (word *length* and *first letter*), and visual (font *color* and the onscreen *location* of each word). We used two control conditions as a baseline; in these control conditions, all of the lists were sorted randomly, but we manipulated the presence or absence of the visual features. In two conditions, we manipulated whether the words' appearances were fixed or variable within each list. In six conditions, we asked participants to first study and recall eight lists whose items were sorted by a target feature (e.g., word category), and then study and recall an additional eight lists whose items had the same features but were sorted in a random temporal order. We were interested in how these manipulations might affect participants' recall behaviors on early (manipulated) lists, as well as how order manipulations on early lists would affect recall behaviors on later (randomly ordered) lists. Finally, in an *adaptive* experimental condition, we used participants' recall behaviors on prior lists to manipulate, in real time, the presentation orders of subsequent lists. In this adaptive condition, we varied whether the order in which items were presented agreed or disagreed with how each participant preferred to organize their memories of the studied items.

From a theoretical perspective, we are interested in several core questions organized around the central theme of how structure in our experiences affects how we remember those experiences, as well as how we remember *future* experiences (which may or may not exhibit similar structure). For example, when we distill participants' experiences down to simple word lists that vary (meaningfully) along just a few feature dimensions, are there important differences in these dimensions' influence on participants' memories? Or are all features essentially "equally" influential? Further, are there differences in how specific features influence participants' memories for ongoing versus future experiences? Are there interaction effects between different features, or is the influence of each feature independent of all others'? And are there individual differences in how people organize their memories, or in how participants are influenced by our experimental manipulations? If so, what are those differences and which aspects of memory do they affect?

Materials and methods

Participants

We enrolled a total of 491 members of the Dartmouth College community across 11 experimental conditions. The conditions included two controls (feature-rich and reduced), two visual manipulation conditions [reduced (early) and reduced (late)], six order manipulation conditions (category, size, length, first letter, color, and location), and a final adaptive condition. Each of these conditions is described in the *Experimental design* subsection below.

Participants received either course credit or a one-time $10 cash payment for enrolling in our study. We asked each participant to fill out a demographic survey that included questions about their age, gender, ethnicity, race, education, vision, reading impairments, medications and recent injuries, coffee consumption on the day of testing, and level of alertness at the time of testing. All components of the demographics survey were optional.

One participant elected not to fill out any part of the demographic survey, and all other participants answered some or all of the survey questions.

We aimed to run (to completion) at least 60 participants in each of the two primary control conditions and in the adaptive condition. In all of the other conditions, we set a target enrollment of at least 30 participants. Because our data collection procedures entailed the coordinated efforts of 12 researchers and multiple testing rooms and computers, it was not feasible for individual experimenters to know how many participants had been run in each experimental condition until the relevant databases were synchronized at the end of each working day. We also over-enrolled participants for each condition to help ensure that we met our minimum enrollment targets even if some participants dropped out of the study prematurely or did not show up for their testing session. This led us to exceed our target enrollments for several conditions. Nevertheless, we analyze all viable data in the present paper.

Participants were assigned to experimental conditions based loosely on their date of participation. (This aspect of our procedure helped us to more easily synchronize the experiment databases across multiple testing computers.) Of the 490 participants who opted to fill out the demographics survey, reported ages ranged from 17 to 31 years (mean: 19.1 years; standard deviation: 1.356 years). A total of 318 participants reported their gender as female, 170 reported their gender as male, and two participants declined to report their gender. A total of 442 participants reported their ethnicity as "not Hispanic or Latino," 39 reported their ethnicity as "Hispanic or Latino," and nine declined to report their ethnicity. Participants reported their races as White (345 participants), Asian (120 participants), Black or African American (31 participants), American Indian or Alaska Native (11 participants), Native Hawaiian or Other Pacific Islander (four participants), Mixed race (three participants), Middle Eastern (one participant), and Arab (one participant). A total of five participants declined to report their race. We note that several participants reported more than one of the above racial categories. Participants reported their highest degrees achieved as “Some college” (359 participants), “High school graduate” (117 participants), “College graduate” (seven participants), “Some high school” (five participants), “Doctorate” (one participant), and “Master’s degree” (one participant). A total of 482 participants reported no reading impairments; eight reported having mild reading impairments. A total of 489 participants reported having normal color vision and one participant reported having impaired color vision. A total of 482 participants reported taking no prescription medications and having no recent injuries; four participants reported having ADHD, one reported having dyslexia, one reported having allergies, one reported a recently torn ACL/MCL, and one reported a concussion from several months prior. The participants reported having consumed 0–3 cups of coffee on the day of the testing session (mean: 0.32 cups; standard deviation: 0.58 cups). Participants reported their current level of alertness, and we converted their responses to numerical scores as follows: “very sluggish” (-2), “a little sluggish” (-1), “neutral” (0), “a little alert” (1), and “very alert” (2). Across all participants, the full range of alertness levels were reported (range: -2–2; mean: 0.35; standard deviation: 0.89).

We dropped from our dataset the one participant who reported having abnormal color vision, as well as 38 participants whose data were corrupted due to technical failures while running the experiment or during the daily database merges. In total, this left usable data from 452 participants, broken down by experimental condition as follows: feature-rich (67 participants), reduced (61 participants), reduced (early) (42 participants), reduced (late) (41 participants), category (30 participants), size (30 participants), length (30 participants), first letter (30 participants), color (31 participants), location (30 participants), and adaptive (60 participants). The participant who declined to fill out their demographic survey participated in the location condition, and we verified verbally that they had normal color vision and no significant reading impairments.

Experimental design

Our experiment is a variant of the classic free recall paradigm that we term "*feature-rich free recall*." In feature-rich free recall, participants study 16 lists, each comprised of 16 words that vary along a number of stimulus dimensions (Fig. 1). The stimulus dimensions include two semantic features related to the *meanings* of the words (semantic category, referent object size), two lexicographic features related to the *letters* that make up the words (word length in number of letters, identity of the word's first letter), and two visual features that are independent of the words themselves (font color, presentation location). Each list contains four words from each of four different semantic categories, with two referent object sizes reflected across all of the words. After studying each list, the participant attempts to recall as many words as they can from that list, in any order they choose. Because each individual word is associated with several well-defined (and quantifiable) features, and because each list incorporates a diverse mix of feature values along each dimension, this allows us to estimate which features participants are considering or leveraging in organizing their memories.

Stimuli

The stimuli in our paradigm were 256 English words selected in a previous study (Ziman et al., 2018). All words referred to concrete nouns and were chosen from 15 unique semantic categories: body parts, building-related, cities, clothing, countries, flowers, fruits, insects, instruments, kitchen-related, mammals, (US) states, tools, trees, and vegetables. We also tagged each word according to the approximate size of the object it referred to. Words

Figure 1: Feature-rich free recall. After studying lists comprised of words that vary along several feature dimensions, participants verbally recall words in any order (microphone icon). Each experimental condition manipulates word features and/or presentation orders within and/or across lists. The rows display representative (illustrated) examples of items from the first list participants might encounter in each condition. The rectangles during the "Presentation phase" show illustrated screen captures during a series of word presentations. Each word appeared onscreen for 2 seconds, followed by 2 seconds of blank screen. The red microphone icons during the "Recall" phase denote the one-minute verbal recall interval. The labels on the right (and corresponding groupings on the left) denote experimental condition labels.

were labeled as "small" if the referent object was likely able to "fit in a standard shoebox" or "large" if the object was larger than a shoebox. Most semantic categories comprised words that reflected both "small" and "large" object sizes, but several included only one or the other (e.g., all countries, US states, and cities are larger than a shoebox; mean number of different sizes per category: 1.33; standard deviation: 0.49). The number of words in each semantic category also varied from 12–28 (mean number of words per category: 17.07; standard deviation: 4.65). We also identified lexicographic features for each word, including its first letter and length (i.e., number of letters). Across all categories, all possible first letters were represented except for 'Q' (average number of unique first letters per category: 11; standard deviation: 2 letters). Word lengths ranged from 3–12 letters (average: 6.17 letters; standard deviation: 2.06 letters).

We assigned the categorized words into a total of 16 lists with several constraints. First, we required that each list contain exactly four unique words from each of four unique categories. Second, we required that each list contain at least one word representing each of the two object sizes ("small" and "large"). On average, each category was represented in 4.27 lists (standard deviation: 1.16 lists). Aside from these two constraints, we randomly assigned each word to a single list (i.e., such that no words appeared in multiple lists or were omitted entirely). After random assignment, each list contained words with an average of 11.13 unique starting letters (standard deviation: 1.15 letters) and an average length of 6.17 letters (standard deviation: 0.34 letters).

The above assignments of words to lists was performed once across all participants, such that every participant studied the same set of 16 lists. In every condition, we randomized the study order of these lists across participants. For participants in most conditions, on some or all of the lists, we also randomly varied two additional visual features associated with each word: the presentation font color and the word's onscreen location. These attributes were assigned independently for each word (and for every participant). These visual features were varied for words in all lists and conditions except for the "reduced" condition (all lists), the first eight lists of the "reduced (early)" condition, and the last eight lists of the "reduced (late)" condition. In these latter cases, all words were presented in black at the center of the experimental computer's display.

To select a random font color for each word, we drew three integers uniformly and at random from the interval [0, 254], corresponding to the red (r), green (g), and blue (b) color channels for that word. To assign random presentation locations to each word, we selected two floating point numbers uniformly and at random (one for the word's horizontal x-coordinate and the other for its vertical y-coordinate). The bounds of these coordinates were selected to cover the entire visible area of the display without cutting off any part of the words. The words were shown on 27-in (diagonal) Retina 5K iMac displays (resolution: 5120×2880 pixels).

Most of the experimental manipulations we carried out entailed presenting or sorting the presented words differently on the first eight lists participants studied (which we call "*early*" lists) versus on the final eight lists they studied ("*late*" lists). Since every participant studied exactly 16 lists, every list was either "early" or "late" depending on its order in the list study sequence.

Real-time speech-to-text processing

Our experimental paradigm incorporates the Google Cloud Speech API speech-to-text engine (Halpern et al., 2016) to automatically transcribe participants' verbal recalls into text. This allows recalls to be transcribed in real time—a distinguishing feature of the experiment; in typical verbal recall experiments, the audio data must be parsed and transcribed manually. In prior work, we used a similar experimental setup (equivalent to the "reduced" condition in the present study) to verify that the automatically transcribed recalls were sufficiently close to human-transcribed recalls to yield reliable data (Ziman et al., 2018). This real-time speech processing component of the paradigm plays an important role in the "adaptive" condition of the experiment, as described below.

Random conditions (Fig. 1, top four rows)

We used two "control" conditions to evaluate and explore participants' baseline behaviors. We also used performance in these control conditions to help interpret performance in other "manipulation" conditions. In the first control condition, which we call the *feature-rich* condition, we randomly shuffled the presentation order (independently for each participant) of the words on each list. In the second control condition, which we call the *reduced* condition, we randomized word presentations as in the feature-rich condition. However, rather than assigning each word a random color and location, we instead displayed all of the words in black and at the center of the screen.

We also designed two conditions in which we varied the words' visual appearances across lists. In the *reduced (early)* condition, we followed the "reduced" procedure (presenting each word in black at the center of the screen) for early lists, and followed the "feature-rich" procedure (presenting each word in a random color and location) for late lists. Finally, in the *reduced (late)* condition, we followed the feature-rich procedure for early lists and the reduced procedure for late lists.

Order manipulation conditions (Fig. 1, middle six rows)

Each of six *order manipulation* conditions used a different feature-based sorting procedure to order words on early lists, where each sorting procedure relied on one relevant feature dimension. All of the irrelevant features varied freely across words on early lists, in that we did not consider irrelevant features in ordering the early lists. However, we note that some features were correlated—for example, some semantic categories of words referred to objects that tended to be a particular size, which meant that category and size were not fully independent (Fig. S9). On late lists, the words were always presented in a randomized order (chosen anew for each participant). In all of the order manipulation conditions, we varied words' font colors and onscreen locations as in the feature-rich condition.

Defining feature-based distances. Sorting words according to a given relevant feature requires first defining a distance function for quantifying the dissimilarity between the values of that feature for each pair of words. This function varied according to the type of feature under consideration. Semantic features (category and size) are *categorical*. For these features, we defined a binary distance function: two words were considered to "match" (i.e., have a distance of 0) if their labels were the same (i.e., both from the same semantic category or both of the same size). If two words' labels were different for a given feature, we defined the words to have a distance of 1. Lexicographic features (length and first letter) are *discrete*. For these features, we defined a discrete distance function. Specifically, we defined the distance between two words as either the absolute difference between their lengths, or the absolute distance between their starting letters in the English alphabet, respectively. For example, two words that started with the same letter would have a "first letter" distance of 0, and a pair of words starting with 'J' and 'A' would have a first letter distance of 9. Because words' lengths and letters' positions in the alphabet are always integers, these discrete distances always take on integer values. Finally, the visual features (color and location) are *continuous* and *multivariate*, in that each "feature" is defined by multiple (positive) real values. We defined the "color" and "location" distances between two words as the Euclidean distances between their (r, g, b) color vectors and (x, y) location vectors (specified as percentages of screen width and height), respectively. Therefore, the color and location distance measures always take on non-negative real values (upper-bounded at 439.94 for color, or 124.52 for location, reflecting the distances between the corresponding maximally different vectors).

Constructing feature-sorted lists. Given a list of words, a relevant feature, and each word's value(s) for that feature, we developed a stochastic algorithm for (noisily) sorting the words. The stochastic aspect of our sorting procedure enabled us to obtain unique orderings for each participant. First, we choose a word uniformly and at random from the set of words on the to-be-presented list. Second, we compute the distances between the chosen word's feature(s) and the corresponding feature(s) of all yet-to-be-presented words. Third, we convert these distances (between the previously presented word's feature values, a, and the candidate word's feature values, b) to similarity scores:

$$\mathrm{similarity}(a, b) = \exp\{-\tau \cdot \mathrm{distance}(a, b)\}, \tag{1}$$

where $\tau = 1$ in our implementation. We note that increasing the value of τ would amplify the influence of similarity on order, and decreasing the value of τ would diminish the influence of similarity on order. Also note that this approach requires $\tau > 0$. Finally, we compute a set of normalized similarity values by dividing the similarities by their sum:

$$\mathrm{similarity}_{\mathrm{normalized}}(a, b) = \frac{\mathrm{similarity}(a, b)}{\sum_{i=1}^{n} \mathrm{similarity}(a, i)}, \tag{2}$$

where in the denominator, i takes on the feature value of each of the n to-be-presented words. The resulting set of normalized similarity scores sums to 1.

As illustrated in Figure 2, we use these normalized similarity scores to construct a sequence of "sticks" that we lay end to end in a line. Each of the n sticks corresponds to a single to-be-presented word, and the stick lengths are proportional to the relative

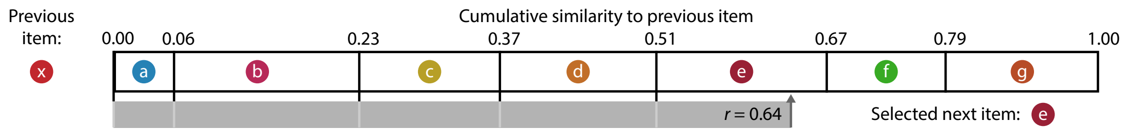

Figure 2: Generating stochastic feature-sorted lists. For a given feature dimension (e.g., color), we compute the similarity (Eqn. 1) between the feature value(s) of the previous item, x, and all yet-to-be-presented items (a–g). Next, we normalize these similarity scores so that they sum to 1. We lay, in sequence, a set of "sticks," one for each candidate item, whose lengths are equal to these normalized similarity scores. To select the next to-be-presented item, we draw a random number r from the uniform distribution bounded between 0 and 1 (inclusive). The identity of the next item is given by the stick adjacent to an indicator that moves distance r (starting from 0) along the sequence of sticks. In this case, the next to-be-presented item is e. Note that each item's chances of selection is proportional to its similarity to the previous item, along the given feature dimension (e.g., color).

similarities between each word's feature value(s) and the feature value(s) of the just-presented word. We choose the next to-be-presented word by moving an indicator along the set of sticks, by a distance chosen uniformly and at random on the interval $[0, 1]$. We select the word associated with the stick lying next to the indicator to be presented next. This process continues iteratively (re-computing the similarity scores and stochastically choosing the next to-be-presented word using the just-presented word) until all of the words have been presented. The result is an ordered list that tends to change gradually along the selected feature dimension (for examples of "sorted" lists, see Fig. 1, *Order manipulation* lists).

Adaptive condition

We designed the *adaptive* experimental condition to study the effect on memory of lists that matched (or mismatched) the ways participants "naturally" organized their memories. Like the other conditions, all participants in the adaptive condition studied a total of 16 word lists in a randomized order. We varied the words' colors and locations for every word presentation, as in the feature-rich and order manipulation conditions.

All participants in the adaptive condition began the experiment by studying a set of four *initialization* lists. Words on these lists were presented in a randomized order (computed independently for each participant). These initialization lists were used to estimate each participant's "memory fingerprint," which we define below. At a high level, a participant's memory fingerprint describes how they prioritize or consider different semantic, lexicographic, and/or visual features when they organize their memories.

Next, participants studied a sequence of 12 lists in three batches of four lists each. These batches came in three types: *random*, *stabilize*, and *destabilize*. The batch types determined how words on the lists in that batch were ordered. Lists in each batch were always presented consecutively (e.g., a participant might receive four random lists, followed by four stabilize lists, followed by four destabilize lists). The batch orders were evenly counterbalanced across participants: there are six possible orderings of the three batches, and 10 participants were randomly assigned to each ordering sub-condition.

Lists in the random batches were sorted randomly (as on the initialization lists and in the feature-rich condition). Lists in the stabilize and destabilize batches were sorted in ways that either matched or mismatched each participant's memory fingerprint, respectively. Our procedures for estimating participants' memory fingerprints and ordering the stabilize and destabilize lists are described next.

Feature clustering scores (uncorrected). Feature clustering scores describe participants' tendencies to recall similar presented items together in their recall sequences, where "similarity" considers one given feature dimension (e.g., category, color, etc.). We based our main approach to computing clustering scores on analogous temporal and semantic clustering scores developed by Polyn et al. (2009). Computing the clustering score for one feature dimension starts by considering the corresponding feature values from the first word the participant recalled correctly from the just-studied list. Next, we sort all not-yet-recalled words in ascending order according to their feature-based distance to the just-recalled item (see *Defining feature-based distances*). We then compute the percentile rank of the observed next recall. We average these percentile ranks across all of the participant's recalls for the current list to obtain a single uncorrected clustering score for the list, for the given feature dimension. We repeated this process for each feature dimension in turn to obtain a single uncorrected clustering score for each list, for each feature dimension.

Temporal clustering score (uncorrected). Temporal clustering describes a participant's tendency to organize their recall sequences by the learned items' encoding positions. For instance, if a participant recalled the lists' words in the exact order they were presented (or in exact reverse order), this would yield a score of 1. If a participant recalled the words in a random order, this would yield an expected score of 0.5. For each recall transition (and separately for each participant), we sorted all not-yet-recalled words according to their absolute lag (i.e., their distance from the just-recalled word in the presented list). We then computed the percentile rank of the next word the participant recalled. We took an average of these percentile ranks across all of the participant's recalls to obtain a single (uncorrected) temporal clustering score for the participant.

Permutation-corrected feature clustering scores. Suppose that two lists contain unequal numbers of items of each size. For example, suppose that list A contains all "large" items, whereas list B contains an equal mix of "large" and "small" items. For a participant recalling list A, any correctly recalled item will necessarily match the size of the previous correctly recalled item. In other words, successively recalling several list A items of the same size is essentially meaningless, since *any* correctly recalled list A word will be large. In contrast, successively recalling several list B items of the same size *could* be meaningful, since (early in the recall sequence) the yet-to-be-recalled items come from a mix of sizes.

However, once all of the small items on list B have been recalled, the best possible next matching recall will be a large item. All subsequent correct recalls must also be large items—so for those later recalls it becomes difficult to determine whether the participant is successively recalling large items because they are organizing their memories according to size, or (alternatively), whether they are simply recalling the yet-to-be-recalled items in a random order. In general, the precise order and blend of feature values expressed in a given list, the order and number of correct recalls a participant makes, the number of intervening presentation positions between successive recalls, and so on, can all affect the range of clustering scores that are possible to observe for a given list. An uncorrected clustering score therefore conflates participants' actual memory organization with other "nuisance" factors.

Following our prior work (Heusser et al., 2017), we used a permutation-based correction procedure to help isolate the behavioral aspects of clustering that we were most interested in. After computing the uncorrected clustering score (for the given list and observed recall sequence), we constructed a "null" distribution of n additional clustering scores by repeatedly randomly shuffling the order of the recalled words and recomputing the clustering score for these shuffled recall sequences (we use $n = 500$ in the present study). This null distribution represents an approximation of the range of clustering scores one might expect to observe by "chance," given that a hypothetical participant was *not* truly clustering their recalls, but where the hypothetical participant still studied and recalled exactly the same items (with the same features) as the true participant. We define the *permutation-corrected clustering score* as the percentile rank of the observed uncorrected clustering score in this estimated null distribution. In this way, a corrected score of 1 indicates that the observed score was greater than any clustering score one might expect by chance—in other words, good evidence that the participant was truly clustering their recalls along the given feature dimension. We applied this correction procedure to all of the clustering scores (feature and temporal) reported in this paper. In Figure S4, we report how participants' clustering scores along different feature dimensions (in the order manipulation conditions) are correlated, and how clustering scores change across lists.

Memory fingerprints. We define each participant's *memory fingerprint* as the set of their permutation-corrected clustering scores across all dimensions we tracked in our study, including their six feature-based clustering scores (category, size, length, first letter, color, and location) and their temporal clustering score. Conceptually, a participant's memory fingerprint describes their tendency to order in their recall sequences (and, presumably, organize in memory) the studied words along each dimension. To obtain stable estimates of these fingerprints for each participant, we averaged their clustering scores across lists. We also tracked and characterized how participants' fingerprints changed across lists (e.g., Figs. 6, S8).

Online "fingerprint" analysis. The presentation orders of some lists in the adaptive condition of our experiment (see *Adaptive condition*) were sorted according to each individual participant's *current* memory fingerprint, estimated using all of the lists they had studied up to that point in the experiment. Because our experiment incorporated a speech-to-text component, all of the behavioral data for each participant could be analyzed just a few seconds after the conclusion of the recall intervals for each list. We used the `Quail` Python package (Heusser et al., 2017) to apply speech-to-text algorithms to the just-collected audio data, aggregate the data for the given participant, and estimate the participant's memory fingerprint using all of their available data up to that point in the experiment. Two aspects of our implementation are worth noting: First, because memory fingerprints are computed independently for each list and then averaged across lists, the already-computed memory fingerprints for earlier lists could be cached and retrieved as needed in future computations. This meant that updating our estimate of a participant's memory fingerprint required computing feature and temporal clustering scores only for the single most recent list. Second, the clustering scores for each dimension of a participant's memory fingerprint could be estimated independently from the others, as could each element of the null distributions of uncorrected clustering scores computed for each dimension (see *Permutation-corrected feature clustering scores*). This enabled us to aggressively parallelize the fingerprint-updating procedure and compress the relevant computations into just a few seconds of computing time. The combined processing time for the speech-to-text algorithm, fingerprint computations, and permutation-based ordering procedure (described next) easily fit within the inter-list intervals, where participants paused for a self-paced break before moving on to study and recall the next list.

Ordering "stabilize" and "destabilize" lists by an estimated fingerprint. In the adaptive condition of our experiment, the presentation orders for *stabilize* and *destabilize* lists were chosen to either maximally or minimally (respectively) comport with participants' memory fingerprints. Given a participant's memory fingerprint and a to-be-presented set of items, we designed a permutation-based procedure for ordering the items. First, we dropped from the participant's fingerprint the temporal clustering score. For the remaining feature dimensions, we arranged the clustering scores in the fingerprint into a template vector f. Second, we computed $n = 2500$ random permutations of the to-be-presented items. These permutations served as candidate presentation orders. We sought to select the specific order that most (or least) closely matched f. Third, for each random permutation, we computed the (permutation-corrected) "fingerprint," treating the permutation as though it were a potential "perfect" recall sequence. (We did not include temporal clustering scores in these fingerprints, since the temporal clustering score for every permutation is always equal to 1.) This yielded a "simulated fingerprint" vector $\hat{f}_p$ for each permutation p. We used these simulated fingerprints to select a specific permutation i that either maximized (for stabilize lists) or minimized (for destabilize lists) the correlation between $\hat{f}_i$ and f.

Computing low-dimensional embeddings of memory fingerprints

Following some of our prior work (Fitzpatrick et al., 2023; Heusser et al., 2021, 2018; Manning et al., 2022), we used low-dimensional embeddings to help visualize how participants' memory fingerprints change across lists (Figs. 6A, S8A). To compute a shared embedding space across participants and experimental conditions, we concatenated the full set of across-participant average fingerprints (for all lists and experimental conditions) to create a large matrix with number-of-lists (16) $\times$ number-of-conditions (10, including the adaptive condition) rows and seven columns (one for each feature clustering score, plus an additional temporal clustering score column). We used principal components analysis to project the seven-dimensional observations into a two-dimensional space (using the two principal components that explained the most variance in the data). For two visualizations (Figs. 6B, S8B), we computed an additional set of two-dimensional embeddings for the *average* fingerprints across lists within a given list grouping (i.e., early or late). For those visualizations, we averaged across the rows (for each condition and group of lists) in the combined fingerprint matrix prior to projecting it into the shared two-dimensional space. This yielded a single two-dimensional coordinate for each *list group* (in each condition), rather than for each individual list. We used these embeddings solely for visualization; all statistical tests were carried out in the original (seven-dimensional) feature spaces.

Factoring out the effects of temporal clustering

For a given list of words, if the values along two feature dimensions (e.g., category and size) are correlated, then the clustering scores for those two dimensions will also be correlated. When lists are sorted along a given feature dimension, the sorted feature values will also tend to be correlated with the serial positions of the words in the list. This means that the temporal clustering score will *also* tend to be correlated with the clustering scores for the sorted feature dimension. These correlations mean that it can be difficult to specifically identify when participants are using one feature versus another (or a manipulated feature versus temporal information) to organize or search their memories.

We developed a permutation-based procedure to factor out the effects of temporal clustering from the clustering scores for each feature dimension. For a given set of recalled items (whose presentation positions are given by $x_1, x_2, x_3, \ldots, x_N$), we circularly shifted the presentation positions by a randomly chosen amount (between 1 and the list length) to obtain a new set of items at the (now altered) positions of the original recalls. Since the new set of items will have the same (average) temporal distances between successive recalls, the temporal clustering score for the new set of items will be equal (on average) to the temporal clustering score for the original recalls. However, we can then re-compute the feature clustering score for those new items. Finally, we can compute a "temporally corrected" feature clustering score by computing the average percentile rank of the observed (raw) feature clustering score within the distributions of circularly shifted feature clustering scores, across $N = 500$ repetitions of this procedure. This new temporally corrected score provides an estimate of the observed degree of feature clustering over and above what could be accounted for by temporal clustering alone.

While these temporally corrected clustering scores are useful for identifying when feature clustering cannot be accounted for by temporal clustering alone, they are *not* necessarily valid estimates of the "true" degree to which participants are organizing their memories along a given feature dimension. For example, on a list where the presentation order and feature values (along the given feature dimension) are perfectly correlated, the temporally corrected score will have an expected value of 0.5 no matter which words a participant recalls, or the order in which they recall them. Therefore these temporally corrected clustering scores are interpretable only to the extent that presentation order and feature value are decoupled.

Analyses

Probability of n^{th} recall curves

Probability of first recall curves (Atkinson and Shiffrin, 1968; Postman and Phillips, 1965; Welch and Burnett, 1924) reflect the probability that an item will be recalled first, as a function of its serial position during encoding. We used an analogous approach to compute the proportion of trials on which each item (as a function of its presentation position) was recalled at each output position n (Hogan, 1975; Howard and Kahana, 1999; Polyn et al., 2009; Zhang et al., 2023). To carry out this analysis, we initialized (for each participant) a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then, for each list, we found the presentation index of the word that was recalled first, and filled in that position in the matrix with a 1. Finally, we averaged over the rows of the matrix to obtain a 1 by 16 array of probabilities, for each participant. We used an analogous procedure to compute probability of n^{th} recall curves for each participant. Specifically, we filled in the corresponding matrices according to the n^{th} recall on each list that each participant made. When a given participant had made fewer than n recalls for a given list, we simply excluded that list from our analysis when computing that participant's curve(s). The probability of first recall curve corresponds to a special case where $n = 1$.

We note that several other studies have used a slightly different approach to compute these curves, by correcting for the "availability" of a given word to be recalled. For example, if a participant recalls item 1, then item 2 on a given list, our approach places a 0 into the item 1 column for that list when computing the "probability of second recall" curve. However, accounting for the fact that the participant had already recalled item 1, an alternative approach (e.g., Farrell, 2010) would be to count the item 1 column as "unobserved" (i.e., missing data). Ultimately we chose to use the simpler variant of this approach in our work, but we direct the reader to further discussion of this issue in other work (Farrell, 2014; Moran and Goshen-Gottstein, 2014).

Lag-conditional response probability curve

The lag-conditional response probability (lag-CRP) curve (Kahana, 1996) reflects the probability of recalling a given item after the just-recalled item, as a function of the items' relative encoding positions (lag). In other words, a lag of 1 indicates that a recalled item was presented immediately after the previously recalled item, and a lag of -3 indicates that a recalled item came three items before the previously recalled item. For each recall transition (following the first recall), we computed the lag between the presentation positions of the just-recalled word and the next-recalled word. We then computed the proportions of transitions (between successively recalled words) for each lag, normalizing for the total numbers of possible transitions. In carrying out this analysis, we excluded all incorrect recalls and repetitions (i.e., recalling a word that had already appeared in the current recall sequence). This yielded, for each list, a 1 by number-of-lags (-15 to $+15$; 30 lags in total, excluding lags of 0) array of conditional probabilities. We averaged these probabilities across lists to obtain a single lag-CRP for each participant. Because transitions at large absolute lags are rare, these curves are typically displayed using range restrictions (Kahana, 2012).

Serial position curve

Serial position curves (Murdock, 1962) reflect the proportion of participants who remember each item as a function of the items' serial positions during encoding. For each participant, we initialized a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then, for each correct recall, we identified the presentation position of the word and entered a 1 into that position (row: list; column: presentation position) in the matrix. This resulted in a matrix whose entries indicated whether or not the words presented at each position, on each list, were recalled by the participant (depending on whether the corresponding entires were set to 1 or 0). Finally, we averaged over the rows of the matrix to yield a 1 by 16 array representing the proportion of words at each position that the participant remembered.

Identifying event boundaries

We used the distances between feature values for successively presented words (see *Defining feature-based distances*) to estimate "event boundaries" where the feature values changed more than usual (DuBrow and Davachi, 2016; Ezzyat and Davachi, 2011; Manning et al., 2016; Radvansky and Copeland, 2006; Swallow et al., 2011, 2009). For each list, for each feature dimension, we computed the distribution of distances between the feature values for successively presented words. We defined event boundaries (e.g., Fig. 3B) as occurring between any successive pair of words whose distances along the given feature dimension were greater than one standard deviation above the mean for that list. Note that, because event boundaries are defined for each feature dimension, each individual list may contain several sets of event boundaries, each at different moments in the presentation sequence (depending on the feature dimension of interest).

Data and code availability

All of the data analyzed in this manuscript, along with all of the code for carrying out the analyses, may be found at https://github.com/ContextLab/FRFR-analyses. Code for running the non-adaptive experimental conditions may be found at https://github.com/ContextLab/efficient-learning-code. Code for running the adaptive experimental condition may be found at https://github.com/ContextLab/adaptiveFR. We have also released an associated Python toolbox for analyzing free recall data, which may be found at https://cdl-quail.readthedocs.io/en/latest.

Results

While holding the set of words (and the assignments of words to lists) constant, we manipulated two aspects of participants' experiences of studying each list. We sought to understand the effects of these manipulations on participants' memories for the studied words. First, we added two additional sources of visual variation to the individual word presentations: font color and onscreen location. Importantly, these visual features were independent of the meaning or semantic content of the words (e.g., word category, size of the referent, etc.) and of the lexicographic properties of the words (e.g., word length, first letter, etc.). We wondered whether this additional word-independent information might facilitate recall (e.g., by providing new or richer potential ways of organizing or retrieving memories of the studied words; Davachi et al., 2003; Drewnowski and Murdock, 1980; Hargreaves et al., 2012; Madan, 2021; Meinhardt et al., 2020; Slamecka and Barlow, 1979; Socher et al., 2009) or impair recall (e.g., by distracting or confusing participants with irrelevant information Lange, 2005; Marsh et al., 2012, 2015; Reinitz et al., 1992).

Second, we manipulated the orders in which words were studied (and how those orderings changed over time). We wondered whether presenting the same list of words with different appearances (e.g., by manipulating font size and onscreen location) or in different orders (e.g., sorted along one feature dimension versus another) might serve to influence how participants organized their memories of the words (e.g., Manning et al., 2015; Polyn and Kahana, 2008). We also wondered whether some order manipulations might be temporally "sticky" by influencing how *future* lists were remembered (e.g., Baddeley, 1968; Darley and Murdock, 1971; Lohnas et al., 2010; Sirotin et al., 2005; Whitely, 1927).

To obtain a clean preliminary estimate of the consequences on memory of randomly varying the font colors and locations of presented words (versus holding the font color fixed at black, and holding the display locations fixed at the center of the display) we compared participants' performance on the *feature-rich* and *reduced* experimental conditions (see *Random conditions*, Fig. S1). In the feature-rich condition the words' colors and locations varied randomly across words, and in the reduced condition words were always presented in black, at the center of the display. Aggregating across all lists for each participant, we found no difference in recall accuracy (i.e., the proportions of correctly recalled words) for feature-rich versus reduced lists ($t(126) = -0.290, p = 0.772$, Cohen's d (d) $= -0.051$, bootstrap estimated 95% confidence interval (CI) $= [-2.387, 1.768]$). However, participants in the feature-rich condition clustered their recalls substantially more along every dimension we examined (temporal clustering: $t(126) = 10.632, p < 0.001, d = 1.882$, CI $= [7.786, 14.386]$; semantic category clustering: $t(126) = 10.148, p < 0.001, d = 1.796$, CI $= [7.324, 13.778]$; size clustering: $t(126) = 12.033, p < 0.001, d = 2.129$, CI $= [9.030, 15.918]$; word length clustering: $t(126) = 10.720, p < 0.001, d = 1.897$, CI $= [7.442, 15.174]$; first letter clustering: $t(126) = 6.679, p < 0.001, d = 1.182$, CI $= [4.490, 9.611]$; see *Permutation-corrected feature clustering scores* for more information about how we quantified each participant's clustering tendencies.) Taken together, these comparisons suggest that adding new features changes how participants organize their memories of studied words, even when those new features are independent of the words themselves and even when the new features vary randomly across words. We found no evidence that those additional uninformative features were distracting (in terms of their impact on memory performance), but they did affect participants' recall dynamics (measured via their clustering scores).

A core assumption of our approach is that each participant organizes their memories in a unique way. We defined each participant's *memory fingerprint* as the set of their permutation-corrected clustering scores across all dimensions we tracked in our study, including their six feature-based clustering scores (category, size, length, first letter, color, and location) and their temporal clustering score. Conceptually, a participant's memory fingerprint describes their tendency to order, in their recall sequences (and, presumably, organize in memory), the studied words along each dimension. If these memory fingerprints are truly unique to each participant, then we would expect that the estimated fingerprints computed for a given participant, on different lists, should be more similar than the estimated fingerprints computed for different participants. We reasoned that the feature-rich condition would provide the best opportunity to test this assumption, since the clustering scores would not be potentially confounded by order manipulations. To test our "unique memory fingerprint" assumption, we compared the similarity (correlation) between the fingerprint from a single list (from one participant) and (a) the average fingerprint from all other lists from the same participant versus (b) the average fingerprint from each other participant (across all of their lists). Repeating this procedure for all lists and participants, we found that participants' fingerprints on a held-out list are reliably more similar to the same participant's fingerprints on other lists than to other participants' fingerprints ($t(70280) = 5.077, p < 0.001, d = 0.162, \mathrm{CI} = [3.086, 6.895]$). This suggests that participants' fingerprints are stable across lists, and that each participant's fingerprint is unique to them.

We next asked whether adding these incidental visual features to later lists (after the participants had already studied impoverished lists), or removing the visual features from later lists (after the participants had already studied visually diverse lists) might affect memory performance. In other words, we sought to test for potential effects of changing the "richness" of participants' experiences over time. All participants studied and recalled a total of 16 lists; we defined *early* lists as the first eight lists and *late* lists as the last eight lists each participant encountered. To help interpret our results, we compared participants' memories on early versus late lists in the above feature-rich and reduced conditions. Participants in both conditions remembered more words on early versus late lists (feature-rich: $t(66) = 4.553, p < 0.001, d = 0.233, \mathrm{CI} = [2.427, 7.262]$; reduced: $t(60) = 2.434, p = 0.018, d = 0.134, \mathrm{CI} = [0.493, 4.910]$). Participants in the feature-rich (but not reduced) conditions exhibited more temporal clustering on early versus late lists (feature-rich: $t(66) = 2.268, p = 0.027, d = 0.181, \mathrm{CI} = [0.437, 4.425]$; reduced: $t(60) = 0.986, p = 0.328, d = 0.061, \mathrm{CI} = [-0.897, 3.348]$). And participants in both conditions tended to exhibit more semantic clustering on early versus late lists (feature-rich, category: $t(66) = 3.684, p < 0.001, d = 0.220, \mathrm{CI} = [1.733, 5.732]$; feature-rich, size: $t(66) = 1.629, p = 0.108, d = 0.100, \mathrm{CI} = [-0.207, 3.905]$; reduced, category: $t(60) = 2.755, p = 0.008, d = 0.177, \mathrm{CI} = [0.761, 5.189]$; reduced, size: $t(60) = 3.081, p = 0.003, d = 0.201, \mathrm{CI} = [1.210, 5.326]$). Participants in the reduced (but not feature-rich) conditions tended to exhibit more lexicographic clustering on early versus late lists (feature-rich, word length: $t(66) = -0.100, p = 0.921, d = -0.010, \mathrm{CI} = [-2.217, 1.899]$; feature rich, first letter: $t(66) = 0.412, p = 0.681, d = 0.045, \mathrm{CI} = [-1.645, 2.461]$; reduced, word length: $t(60) = 3.762, p < 0.001, d = 0.261, \mathrm{CI} = [1.604, 6.821]$; reduced, first letter:

$t(60) = 1.721, p = 0.090, d = 0.175, \mathrm{CI} = [-0.138, 4.098]$). Taken together, these comparisons suggest that even when the presence or absence of incidental visual features is stable across lists, participants still exhibit some differences in their performance and memory organization tendencies for early versus late lists.

With these differences in mind, we next compared participants' memories on early versus late lists for two additional experimental conditions (see *Random conditions*, Fig. S1). In a *reduced (early)* condition, we held the visual features constant on early lists, but allowed them to vary randomly on late lists. In a *reduced (late)* condition, we allowed the visual features to vary randomly on early lists, but held them constant on late lists. Given our above findings that (a) participants tended to exhibit stronger clustering effects on feature-rich (versus reduced) lists, and (b) participants tended to remember more words and exhibit stronger clustering effects on early (versus late) lists, we expected these early versus late differences to be enhanced in the reduced (early) condition and diminished in the reduced (late) condition. However, to our surprise, participants in *neither* condition exhibited reliable early versus late differences in accuracy (reduced (early): $t(41) = 1.499, p = 0.141, d = 0.098, \mathrm{CI} = [-0.345, 3.579]$; reduced (late): $t(40) = 1.462, p = 0.152, d = 0.121, \mathrm{CI} = [-0.376, 2.993]$), temporal clustering (reduced (early): $t(41) = 0.857, p = 0.396, d = 0.068, \mathrm{CI} = [-1.012, 2.896]$; reduced (late): $t(40) = 1.244, p = 0.221, d = 0.128, \mathrm{CI} = [-0.894, 3.088]$), nor feature-based clustering (reduced (early), category: $t(41) = 0.707, p = 0.484, d = 0.068, \mathrm{CI} = [-1.314, 2.830]$; reduced (early), size: $t(41) = 0.803, p = 0.427, d = 0.079, \mathrm{CI} = [-1.142, 2.953]$; reduced (early), length: $t(41) = 0.461, p = 0.648, d = 0.060, \mathrm{CI} = [-1.545, 2.462]$; reduced (early), first letter: $t(41) = 0.781, p = 0.439, d = 0.101, \mathrm{CI} = [-1.039, 2.881]$; reduced (late), category: $t(40) = 0.101, p = 0.920, d = 0.009, \mathrm{CI} = [-1.776, 2.307]$; reduced (late), size: $t(40) = 0.555, p = 0.582, d = 0.058, \mathrm{CI} = [-1.444, 2.274]$; reduced (late), length: $t(40) = 1.482, p =$

$0.146, d = 0.126, \mathrm{CI} = [-0.444, 3.743]$; reduced (late), first letter: $t(40) = -0.143, p = 0.887, d = -0.017, \mathrm{CI} = [-2.204, 1.830]$). We hypothesized that adding or removing the variability in the visual features was acting as a sort of "event boundary" between early and late lists (e.g., Clewett et al., 2019; Radvansky and Copeland, 2006; Radvansky and Zacks, 2017). In prior work, we (and others) have found that memories formed just after event boundaries can be enhanced (e.g., due to less contextual interference between pre- and post-boundary items; Flores et al., 2017; Gold et al., 2017; Manning et al., 2016; Pettijohn et al., 2016).

We found that *adding* incidental visual features on later lists that had not been present on early lists (as in the reduced (early) condition) served to enhance recall performance relative to conditions where all lists had the same blends of features (accuracy for feature-rich versus reduced (early): $t(107) = -2.230, p = 0.028, d = -0.439, \mathrm{CI} = [-4.252, -0.229]$; reduced versus reduced (early): $t(101) = -2.045, p = 0.043, d = -0.410, \mathrm{CI} = [-3.826, 0.112]$; also see Fig. S3A). However, *subtracting* irrelevant visual features on later lists that *had* been present on early lists (as in the reduced (late) condition) did not appear to impact recall performance (accuracy for feature-rich versus reduced (late): $t(106) = -0.638, p = 0.525, d = -0.126, \mathrm{CI} = [-2.720, 1.362]$; reduced versus reduced (late): $t(100) = -0.407, p = 0.685, d = -0.082, \mathrm{CI} = [-2.477, 1.626]$). These comparisons suggest that recall accuracy has a directional component: accuracy is affected differently by removing features later that had been present earlier versus adding features later that had *not* been present earlier. In contrast, we found that participants exhibited more temporal and feature-based clustering when we added incidental visual features to *any* lists (comparisons of clustering on feature-rich versus reduced lists are reported above; temporal clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions: $t\text{s} \leq -9.885, p\text{s} < 0.001$; feature-based clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions:

ts ≤ -4.555, ps < 0.001). Temporal and feature-based clustering were not reliably different in the feature-rich, reduced (early), and reduced (late) conditions (temporal clustering in feature-rich versus reduced (early) and feature-rich versus reduced (late) conditions: ts ≥ -1.379, ps ≥ 0.171; feature-based clustering in feature-rich versus reduced (early) and feature-rich versus reduced (late) conditions: $|t|$s ≤ 1.441, ps ≥ 0.153).

Taken together, our findings thus far suggest that adding item features that change over time, even when they vary randomly and independently of the items, can enhance participants' overall memory performance and can also enhance temporal and feature-based clustering. To the extent that the number of item features that vary from moment to moment approximates the "richness" of participants' experiences, our findings suggest that participants remember "richer" stimuli better and organize richer stimuli more reliably in their memories. Next, we turn to examine the memory effects of varying the temporal ordering of different stimulus features. We hypothesized that changing the orders in which participants were exposed to the words on a given list might enhance (or diminish) the relative influence of different features. For example, presenting a set of words alphabetically might enhance participants' attention to the studied items' first letters, whereas sorting the same list of words by semantic category might instead enhance participants' attention to the words' semantic attributes. Importantly, we expected these order manipulations to hold even when the variation in the total set of features (across words) was held constant across lists (e.g., unlike in the reduced (early) and reduced (late) conditions, where variations in visual features were added or removed from a subset of the lists participants studied).

Across each of six order manipulation conditions, we sorted early lists by one feature dimension but randomly ordered the items on late lists (see *Order manipulation conditions*; features: category, size, length, first letter, color, and location). Participants in the category-

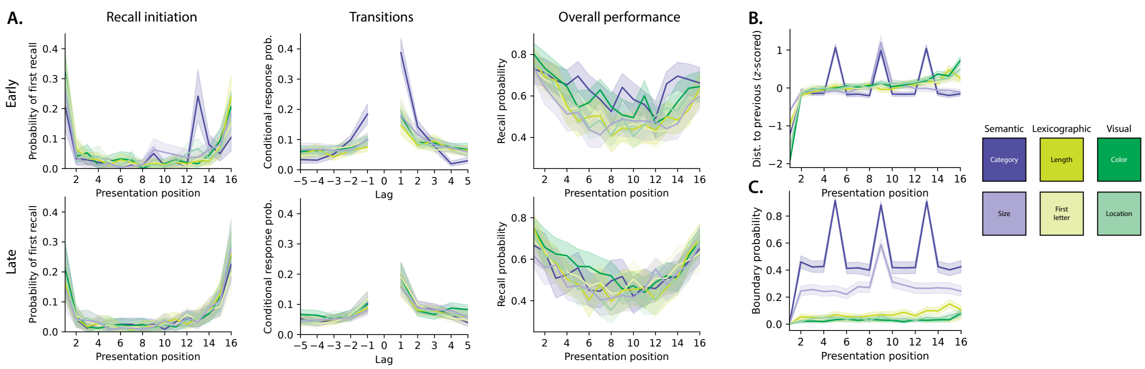

Figure 3: Recall dynamics in feature-rich free recall (order manipulation conditions). A. Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling each word are plotted as a function of the relative position (Lag) to the words recalled just-prior. **Right panels.** The overall probabilities of recalling each word are plotted as a function of presentation position. **All panels.** Error ribbons denote bootstrap-estimated 95% confidence intervals (calculated across participants). Top panels display the recall dynamics for early (order manipulation) lists in each condition (color). Bottom panels display the recall dynamics for late (randomly ordered) lists. See Figures S1 and S2 for analogous plots for the random and adaptive conditions. **B.** Feature distances (z-scored within condition) between the features of successively presented words (see *Defining feature-based distances*), for each condition's feature of focus, plotted as a function of presentation position.**C.** Proportion of event boundaries (see *Identifying event boundaries*) for each condition's feature of focus, plotted as a function of presentation position.

ordered condition showed an increase in memory performance on early lists (accuracy, relative to early feature-rich lists; $t(95) = 3.034, p = 0.003, d = 0.667, \mathrm{CI} = [1.048, 5.113]$). Participants in the color-ordered condition also showed a trending increase in memory performance on early lists (again, relative to early feature-rich lists: $t(96) = 1.850, p = 0.067, d = 0.402, \mathrm{CI} = [-0.010, 3.712]$; Fig. 5A). Participants' performances on early lists in all of the other order manipulation conditions were indistinguishable from performance on the early feature-rich lists ($|t|\mathrm{s} \leq 1.013, p\mathrm{s} \geq 0.314$). Participants in both of the semantically ordered conditions exhibited stronger temporal clustering on early lists (versus early feature-rich lists; category: $t(95) = 8.813, p < 0.001, d = 1.936, \mathrm{CI} = [6.793, 11.751]$; size: $t(95) = 2.630, p = 0.010, d = 0.578, \mathrm{CI} = [0.831, 4.866]$; Fig. 5B). Participants in the length-ordered condition tended to exhibit *less* temporal clustering on early lists relative to early feature-rich lists ($t(95) = -1.547, p = 0.125, d = -0.340, \mathrm{CI} = [-3.693, 0.341]$), whereas participants in the first letter-ordered condition exhibited stronger temporal clustering on early lists ($t(95) = 2.858, p = 0.005, d = 0.628, \mathrm{CI} = [1.031, 4.886]$). Participants in the visually ordered conditions exhibited more similar performance (accuracy) on early lists, relative to early feature rich lists (we found a trending enhancement for participants in the color-ordered condition: $t(96) = 1.850, p = 0.067, d = 0.402, \mathrm{CI} = [-0.010, 3.712]$; location: $t(95) = 0.043, p = 0.966, d = 0.010, \mathrm{CI} = [-1.598, 1.729]$). Participants in the visually ordered conditions also showed similar temporal clustering on early lists, relative to early feature-rich lists (color: $t(96) = -1.339, p = 0.184, d = -0.291, \mathrm{CI} = [-3.238, 0.394]$, we found a trending increase for participants in the location-ordered condition: $t(95) = 1.705, p = 0.092, d = 0.374, \mathrm{CI} = [-0.155, 3.521]$). We also compared feature-based clustering on early lists across the order manipulation and feature-rich conditions. Since these results were similar across both semantic conditions (category and size), both lexicographic conditions (length and first letter), and both visual conditions (color and location), here we aggregate data from conditions that manipulated each of these three feature groupings in our comparisons, to simplify the presentation. On early lists, participants in the semantically ordered conditions exhibited stronger semantic clustering relative to participants in the feature-rich condition (category: $t(125) = 2.722, p = 0.007, d = 0.484, \mathrm{CI} = [0.827, 4.932]$; size: $t(125) = 3.866, p < 0.001, d = 0.687, \mathrm{CI} = [2.020, 5.983]$), but showed no reliable differences in lexicographic (length: $t(125) = 0.521, p = 0.603, d = 0.093, \mathrm{CI} = [-1.311, 2.333]$; first letter: $t(125) = -0.842, p = 0.401, d = -0.150, \mathrm{CI} = [-2.825, 1.095]$) or visual (color: $t(125) = -0.650, p = 0.517, d = -0.116, \mathrm{CI} = [-2.680, 1.249]$; location: $t(125) = -0.251, p = 0.802, d = -0.045, \mathrm{CI} = [-2.257, 1.524]$) clustering. Similarly, participants in the lexicographically ordered conditions exhibited stronger (relative to feature rich participants) lexicographic clustering (length: $t(125) = 3.682, p < 0.001, d = 0.655, \mathrm{CI} = [1.890, 5.569]$; first letter: $t(125) = 5.134, p < 0.001, d = 0.912, \mathrm{CI} = [3.251, 7.258]$) on early lists, but showed no reliable differences in semantic (category: $t(125) = -1.040, p = 0.301, d = -0.185, \mathrm{CI} = [-3.095, 1.092]$; size: $t(125) = 0.006, p = 0.995, d = 0.001, \mathrm{CI} = [-1.933, 1.952]$) or visual (color: $t(125) = 0.092, p = 0.927, d = 0.016, \mathrm{CI} = [-1.834, 1.867]$; location: $t(125) = 0.407, p = 0.685, d = 0.072, \mathrm{CI} = [-1.655, 2.463]$) clustering. And participants in the visually ordered conditions exhibited stronger visual clustering (again, relative to feature-rich participants, and on early lists; color: $t(126) = 2.022, p = 0.045, d = 0.358, \mathrm{CI} = [0.056, 3.965]$; location: $t(126) = 4.390, p < 0.001, d = 0.777, \mathrm{CI} = [2.730, 6.199]$), but showed no reliable differences in semantic (category: $t(126) = 0.012, p = 0.991, d = 0.002, \mathrm{CI} = [-1.988, 1.871]$; size: $t(126) = -0.104, p = 0.917, d = -0.018, \mathrm{CI} = [-2.166, 1.847]$) or lexicographic (length: $t(126) = 0.592, p = 0.555, d = 0.105, \mathrm{CI} = [-1.361, 2.420]$; first letter: $t(126) = 0.040, p = 0.968, d = 0.007, \mathrm{CI} = [-1.791, 1.863]$) clustering. Taken together, these order manipulation results suggest several broad patterns (Figs. 3A, 4). First, most of the order manipulations we carried out did *not* reliably affect overall recall performance. Second, most of the order manipulations increased participants' tendencies to temporally cluster their recalls. Third, all of the order manipulations enhanced participants' clustering of each condition's target feature (i.e., semantic manipulations enhanced semantic clustering, lexicographic manipulations enhanced lexicographic clustering, and visual manipulations enhanced visual clustering; Fig. 5C) while leaving clustering along other feature dimensions roughly unchanged (i.e., semantic manipulations did not affect lexicographic or visual clustering, and so on). Although it is not possible to fully separate feature versus temporal clustering when considering sorted lists, we used a permutation-based procedure to identify the degree of feature clustering over and above what could be accounted for by temporal clustering alone (see *Factoring out the effects of temporal clustering*). When we carried out this analysis (Fig. 5D), we found that participants exhibited more semantic clustering on semantically sorted lists than on randomly ordered lists, but the effects of the other order manipulations could not reliably be separated from temporal clustering alone (reliable comparisons are reported in the figure).

When we closely examined the sequences of words participants recalled from early order-manipulated lists (Fig. 3A, top panel), we noticed several differences from the dynamics of participants' recalls of randomly ordered lists (Figs. S1, S7). One difference is that participants in the category condition (dark purple curves, Fig. 3) most often initiated recall with the fourth-from-last item (*Recall initiation*, top left panel), whereas participants who recalled randomly ordered lists tended to initiate recall with either the first or last list items (Fig. S1, top left panel). We hypothesized that the participants might be "clumping" their recalls into groups of items that shared category labels. Indeed, when we compared the positions of feature changes in the study sequence (Fig. 3C; see *Identifying event boundaries*) with the positions of items participants recalled first, we noticed a striking correspondence in both semantic conditions. Specifically, on category-ordered lists,

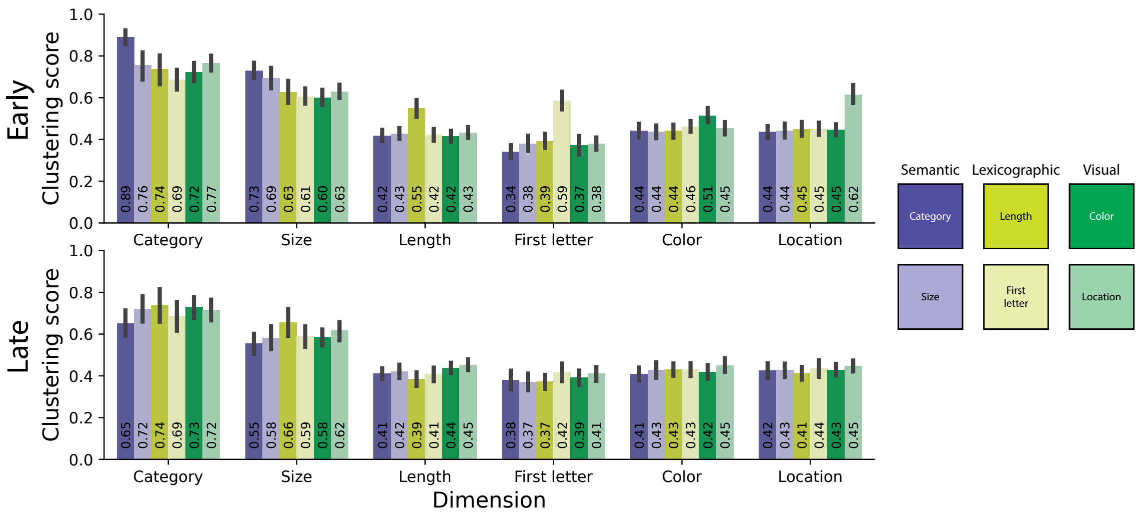

Figure 4: Memory "fingerprints" (order manipulation conditions). The across-participant clustering scores for each feature type (x-coordinate) are displayed for each experimental condition (color), separately for order manipulation (early, top) and randomly ordered (late, bottom) lists. Error bars denote bootstrap-estimated 95% confidence intervals. See Figures S5 and S6 for analogous plots for the random and adaptive conditions.

the category labels changed every four items on average (dark purple peaks in Figs. 3B, C), and participants also seemed to display an increased tendency (relative to other order manipulation and random conditions) to initiate recall of category-ordered lists with items whose study positions were integer multiples of four. Similarly, for size-ordered lists, the size labels changed every eight items on average (light purple peaks in Figs. 3B, C), and participants also seemed to display an increased tendency to initiate recall of size-ordered lists with items whose study positions were integer multiples of eight. A second striking difference is that participants in the category condition exhibited a much steeper lag-CRP (Fig. 3A, top middle panel) than participants in other conditions. (This is another expression of participants' increased tendencies to temporally cluster their recalls on category-ordered lists, as we reported above.) Taken together, these order-specific idiosyncrasies suggest a hierarchical set of influences on participants' memories. At longer timescales, "event boundaries" (to use the term loosely) can be induced across lists by adding or removing incidental visual features. At shorter timescales, "event boundaries" can be induced across items (within a single list) by adjusting how item features change throughout the list.

The above comparisons between memory performance on early lists in the order manipulation versus feature-rich conditions highlight how sorted lists are remembered differently from random lists. We also wondered how sorting lists along each feature dimension influenced memory relative to sorting lists along the other feature dimensions. Participants trended towards remembering early lists that were sorted semantically better than lexicographically sorted lists ($t(118) = 1.936, p = 0.055, d = 0.353, \mathrm{CI} = [0.057, 3.916]$). Participants also remembered visually sorted lists better than lexicographically sorted lists ($t(119) = 2.145, p = 0.034, d = 0.390, \mathrm{CI} = [0.208, 4.254]$). However, participants showed no reliable differences in recall for semantically versus visually sorted lists ($t(119) = 0.113, p = 0.910, d = 0.021, \mathrm{CI} = [-1.987, 2.097]$). Participants temporally clustered semantically sorted lists more strongly than either lexicographically ($t(118) = 5.620, p < 0.001, d = 1.026, \mathrm{CI} = [3.486, 8.010]$) or visually ($t(119) = 6.613, p < 0.001, d = 1.202, \mathrm{CI} = [4.481, 9.464]$) sorted lists, but did not show reliable differences in temporal clustering on lexicographically versus visually sorted lists ($t(119) = 0.589, p = 0.557, d = 0.107, \mathrm{CI} = [-1.336, 2.539]$). Participants also showed reliably more semantic clustering on semantically sorted lists than lexicographically (category: $t(118) = 3.667, p < 0.001, d = 0.670, \mathrm{CI} = [1.822, 5.942]$, size: $t(118) = 3.972, p < 0.001$) or visually (category: $t(119) = 2.702, p = 0.008$, size: $t(118) = 4.043, p < 0.001, d = 0.738, \mathrm{CI} = [2.145, 6.296]$) sorted lists; more lexicographic clustering on lexicographically sorted lists than semantically (length: $t(118) = 3.390, p < 0.001, d = 0.619, \mathrm{CI} = [1.499, 5.661]$; first letter: $t(118) = 5.705, p < 0.001, d = 1.042, \mathrm{CI} = [3.841, 7.790]$) or visually (length: $t(119) =$

$3.399, p < 0.001, d = 0.618, \mathrm{CI} = [1.500, 5.527]$; first letter: $t(119) = 4.859, p < 0.001, d = 0.883, \mathrm{CI} = [2.860, 6.849]$) sorted lists; and more visual clustering on visually sorted lists than semantically (color: $t(119) = 2.673, p = 0.009, d = 0.486, \mathrm{CI} = [0.848, 4.567]$; location: $t(119) = 4.499, p < 0.001, d = 0.818, \mathrm{CI} = [2.721, 6.399]$) or lexicographically (color: $t(119) = 1.988, p = 0.049, d = 0.361, \mathrm{CI} = [0.102, 3.894]$; location: $t(119) = 3.966, p < 0.001, d = 0.721, \mathrm{CI} = [2.099, 5.862]$) sorted lists. In summary, sorting lists by different features appeared to have slightly different effects on overall memory performance and temporal clustering. Participants also tended to cluster their recalls along a given feature dimension more when the studied lists were (versus were not) sorted along that dimension.

Beyond affecting how we process and remember *ongoing* experiences, what is happening to us now can also affect how we process and remember *future* experiences. Within the framework of our study, we wondered: if early lists are sorted along different feature dimensions, might this affect how people remember later (random) lists? In exploring this question, we considered both group-level effects (i.e., effects that tended to be common across individuals) and participant-level effects (i.e., effects that were idiosyncratic across individuals).

At the group level, there seemed to be almost no lingering impact of sorting early lists on memory for later lists. To simplify the presentation, we report these null results in aggregate across the three feature groupings. Relative to memory performance on late feature-rich lists, participants' memory performance in all six order manipulation conditions showed no reliable differences (semantic: $t(125) = 0.487, p = 0.627, d = 0.087, \mathrm{CI} = [-1.661, 2.323]$; lexicographic: $t(125) = 0.878, p = 0.382, d = 0.156, \mathrm{CI} = [-1.226, 3.044]$; visual: $t(126) = 1.437, p = 0.153, d = 0.254, \mathrm{CI} = [-0.447, 3.519]$). Nor did we observe any reliable differences in temporal clustering on late lists (relative to late feature-rich

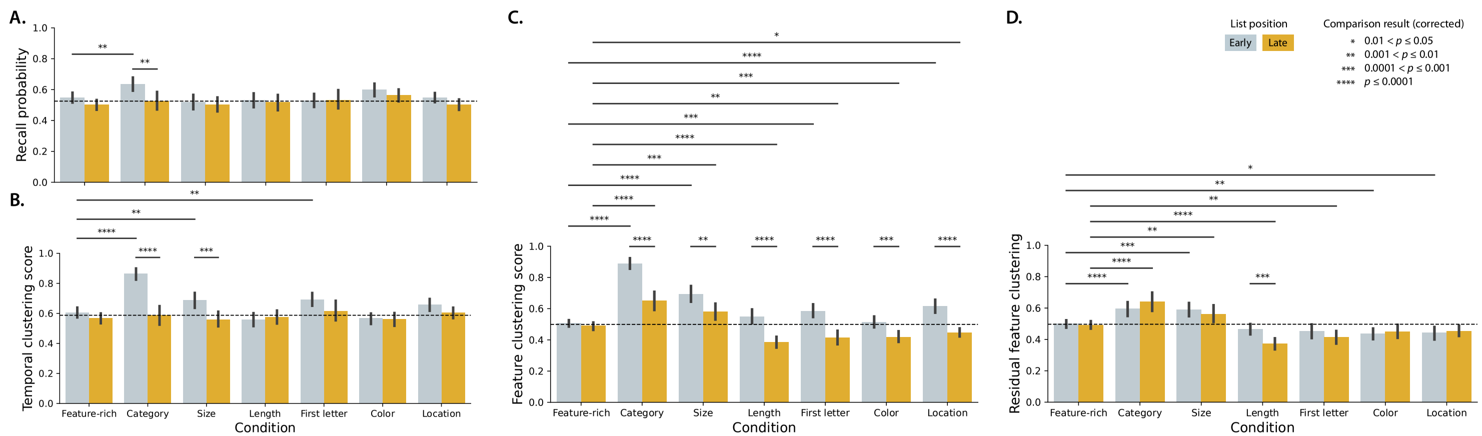

Figure 5: Recall probability and clustering scores on early and late lists. The bar heights display the average (across participants) recall probabilities (**A.**), temporal clustering scores (**B.**), feature clustering scores (**C.**), and residual feature clustering scores (after factoring out temporal clustering effects; **D.**) for early (gray) and late (gold) lists. For the feature-rich bars (left), the feature clustering scores are averaged across all feature dimensions. For the order manipulation conditions, feature clustering scores are displayed for the focused-on feature for each condition (e.g., category clustering scores are displayed for the category condition, and so on). All panels: error bars denote bootstrap-estimated 95% confidence intervals. The horizontal dotted lines denote the average values (across all lists and participants) for the feature rich condition. The bars denote t-tests between the corresponding bars, and the asterisks denote the Benjamini-Hochberg-corrected p-values. Comparisons for which corrected $p \geq 0.05$ are not shown.

lists; semantic: $t(125) = 0.157, p = 0.875, d = 0.028, \mathrm{CI} = [-1.859, 1.974]$; lexicographic: $t(125) = 0.998, p = 0.320, d = 0.177, \mathrm{CI} = [-0.902, 2.920]$; visual: $t(126) = 0.548, p = 0.585, d = 0.097, \mathrm{CI} = [-1.450, 2.365]$). Aside from a slightly increased tendency for participants to cluster words by their length on late visual order manipulation lists (more than late feature-rich lists; $t(126) = 2.005, p = 0.047, d = 0.355, \mathrm{CI} = [0.211, 3.722]$), we observed no reliable differences in any type of feature clustering on late order manipulation condition lists versus late feature-rich lists ($|t|\mathrm{s} \leq 1.124, p\mathrm{s} \geq 0.263$).

We also looked for more subtle group-level patterns. For example, perhaps sorting early lists by one feature dimension could affect how participants cluster *other* features (on early and/or late lists) as well. As described above, a participant's memory fingerprint characterizes how they tend to retrieve memories of the studied items, perhaps searching in parallel through several feature spaces (or along several representational dimensions). To gain insights into the dynamics of how participants' clustering scores tended to change over time, we computed the average (across participants) fingerprint from each list, from each order manipulation condition (Fig. 6). We projected these fingerprints into a two-dimensional space to help visualize the dynamics (top panels; see *Computing low-dimensional embeddings of memory fingerprints*). We found that participants' average fingerprints tended to remain relatively stable on early lists, and exhibited a "jump" to another stable state on later lists. The sizes of these jumps varied somewhat across conditions (the Euclidean distances between fingerprints in their original high dimensional spaces are displayed in the bottom panels). We also averaged the fingerprints across early and late lists, respectively, for each condition (Fig. 6B). We found that participants' fingerprints on early lists seem to be influenced by the order manipulations for those lists (see the locations of the circles in Fig. 6B). There also seemed to be some consistency across different features within a broader type. For example, both semantic

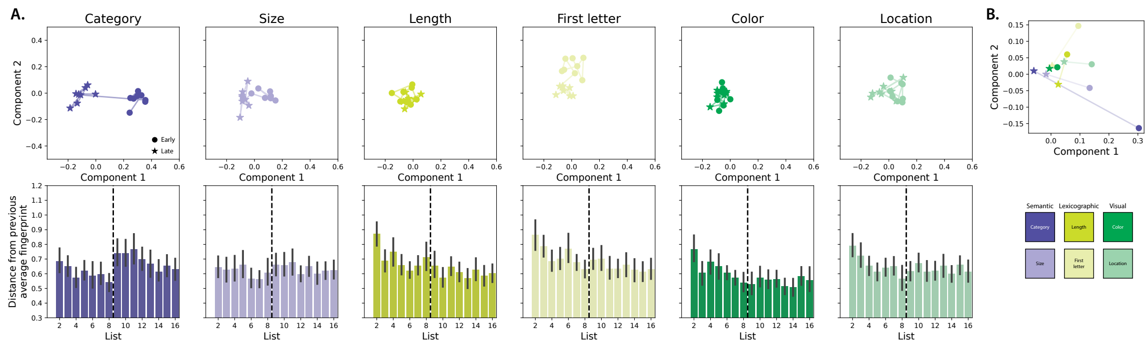

Figure 6: Memory fingerprint dynamics (order manipulation conditions). A. Each column (and color) reflects an experimental condition. In the top panels, each marker displays a 2D projection of the (across-participant) average memory fingerprint for one list. Order manipulation (early) lists are denoted by circles and randomly ordered (late) lists are denoted by stars. All of the fingerprints (across all conditions and lists) are projected into a common space. The bar plots in the bottom panels display the Euclidean distances of the per-list memory fingerprints to the average fingerprint across all prior lists, for each condition. Error bars denote bootstrap-estimated 95% confidence intervals. The dotted vertical lines denote the boundaries between early and late lists. **B.** In this panel, the fingerprints for early (circle) and late (star) lists are averaged across lists and participants before projecting the fingerprints into a (new) 2D space. See Figure S8 for analogous plots for the random conditions.

feature conditions (category and size; purple markers) diverge in a similar direction from the group; both lexicographic feature conditions (length and first letter; yellow markers) diverge in a similar direction; and both visual conditions (color and location; green) also diverge in a similar direction. But on late lists, participants' fingerprints seem to return to a common state that is roughly shared across conditions (i.e., the stars in that panel are clumped together).

When we examined the data at the level of individual participants (Figs. 7 and 8), a clearer story emerged. Within each order manipulation condition, participants exhibited a range of feature clustering scores on both early and late lists (Fig. 7A, B). Across every order manipulation condition, participants who exhibited stronger feature clustering (for their condition's manipulated feature) recalled more words. This trend held overall across conditions and participants (early: $r(179) = 0.492, p < 0.001, \mathrm{CI} = [0.352, 0.606]$;

late: $r(179) = 0.403, p < 0.001, \mathrm{CI} = [0.271, 0.517]$) as well as for each condition individually for early (rs $\geq$ 0.331, all ps $\leq$ 0.069) and late (rs $\geq$ 0.404, all ps $\leq$ 0.027) lists. We found no evidence of a condition-level trend; for example, the conditions where participants tended to show stronger clustering scores were not correlated with the conditions where participants remembered more words (early: $r(4) = 0.511, p = 0.300, \mathrm{CI} = [-0.999, 0.996]$; late: $r(4) = -0.304, p = 0.559, \mathrm{CI} = [-0.833, 0.748]$; see insets of Fig. 7A and B). We observed carryover associations between feature clustering and recall performance (Fig. 7C, D). Participants who showed stronger feature clustering on early lists in the non-visual conditions tended to recall more items on late lists (across conditions: $r(179) = 0.230, p = 0.002, \mathrm{CI} = [0.072, 0.372]$; all non-visual conditions individually: rs $\geq$ 0.405, all ps $\leq$ 0.027; color: $r(29) = 0.212, p = 0.251, \mathrm{CI} = [-0.164, 0.532]$; location: $r(28) = 0.320, p = 0.085, \mathrm{CI} = [0.011, 0.584]$). Participants who recalled more items on early lists also tended to show stronger feature clustering on late lists (across conditions: $r(179) = 0.464, p < 0.001, \mathrm{CI} = [0.321, 0.582]$; individual conditions: all rs $\geq$ 0.377, all ps $\leq$ 0.040). Neither of these effects showed condition-level trends (early feature clustering versus late recall probability: $r(4) = -0.338, p = 0.512, \mathrm{CI} = [-0.971, 0.634]$; early recall probability versus late feature clustering: $r(4) = 0.451, p = 0.369, \mathrm{CI} = [-0.986, 0.998]$). We also looked for associations between feature clustering and temporal clustering. Across every order manipulation condition, participants who exhibited stronger feature clustering also exhibited stronger temporal clustering. For early lists (Fig. 7E), this trend held overall ($r(179) = 0.916, p < 0.001, \mathrm{CI} = [0.893, 0.936]$), for each condition individually (all rs $\geq$ 0.822, all ps < 0.001), and across conditions ($r(4) = 0.964, p = 0.002$). For late lists (Fig. 7F), the results were more variable (overall: $r(179) = 0.348, p < 0.001$; all non-visual conditions: rs $\geq$ 0.382, all ps $\leq$ 0.037; color: $r(29) = 0.453, p = 0.011$; location: $r(28) = 0.190, p = 0.314$; across-conditions: $r(4) = -0.036, p = 0.945$). While less robust than the carryover associations between feature clustering and recall performance, we also observed some carryover associations between feature clustering and temporal clustering (Fig. 7G, H). Participants who showed stronger feature clustering on early lists showed stronger temporal clustering on later lists (overall: $r(179) = 0.464, p < 0.001, \mathrm{CI} = [0.321, 0.582]$; for individual conditions: all rs ≥ 0.377, all ps ≤ 0.040; across conditions: $r(4) = 0.451, p = 0.369, \mathrm{CI} = [-0.986, 0.998]$). And participants who showed stronger temporal clustering on early lists trended towards showing stronger feature clustering on later lists (overall: $r(179) = 0.266, p < 0.001, \mathrm{CI} = [0.129, 0.396]$; for individual conditions: all rs ≥ 0.298, all ps ≤ 0.110; across conditions: $r(4) = 0.064, p = 0.903, \mathrm{CI} = [-0.972,)$. Taken together, the results displayed in Figure 7 show that participants who were more sensitive to the order manipulations (i.e., participants who showed stronger feature clustering for their condition's feature on early lists) remembered more words and showed stronger temporal clustering. These associations also appeared to carry over across lists, even when the items on later lists were presented in a random order.

If participants show different sensitivities to order manipulations, how do their behaviors carry over to later lists? We found that participants who showed strong feature clustering on early lists often tended to show strong feature clustering on late lists (Fig. 8A; overall across participants and conditions: $r(179) = 0.591, p < 0.001, \mathrm{CI} = [0.472, 0.682]$; category: $r(28) = 0.590, p < 0.001, \mathrm{CI} = [0.354, 0.756]$; size: $r(28) = 0.488, p = 0.006, \mathrm{CI} = [0.134, 0.732]$; length: $r(28) = 0.384, p = 0.036, \mathrm{CI} = [0.040, 0.681]$; first letter: $r(28) = 0.202, p = 0.284, \mathrm{CI} = [-0.273, 0.620]$; color: $r(29) = -0.183, p = 0.325, \mathrm{CI} = [-0.562, 0.258]$; location: $r(28) = 0.031, p = 0.870, \mathrm{CI} = [-0.240, 0.296]$; across conditions: $r(4) = 0.942, p = 0.005, \mathrm{CI} = [0.442, 1.000]$). Although participants tended to show weaker feature clustering on late lists (Fig. 6) on *average*, the associations between early and late lists for individual participants suggests that some influence of early order manipulations may linger on late

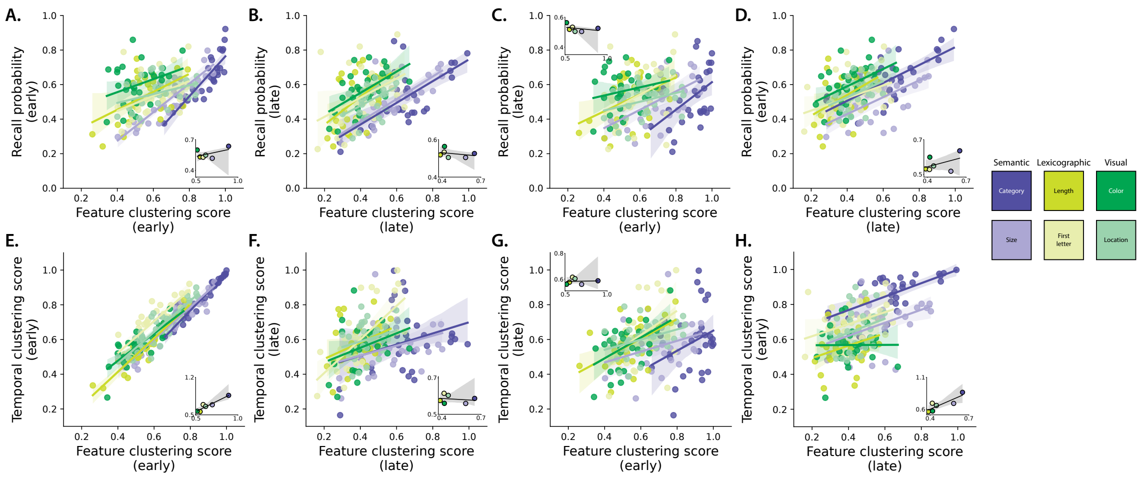

Figure 7: Interactions between feature clustering, recall probability, and contiguity. A. Recall probability versus feature clustering scores for order manipulation (early) lists. **B.** Recall probability versus feature clustering for randomly ordered (late) lists. **C.** Recall probability on late lists versus feature clustering on early lists. **D.** Recall probability on early lists versus feature clustering on late lists. **E.** Temporal clustering scores (contiguity) versus feature clustering scores on early lists. **F.** Temporal clustering scores versus feature clustering scores on late lists. **G.** Temporal clustering scores on late lists versus feature clustering scores on early lists. **H.** Temporal clustering scores on early lists versus feature clustering scores on late lists. **All panels.** Each dot in the main scatterplots denotes the average scores for one participant. The colored regression lines are computed across participants. The inset displays condition-averaged results, where each dot reflects a single condition and the regression line is computed across experimental conditions. All error ribbons denote bootstrap-estimated 95% confidence intervals.

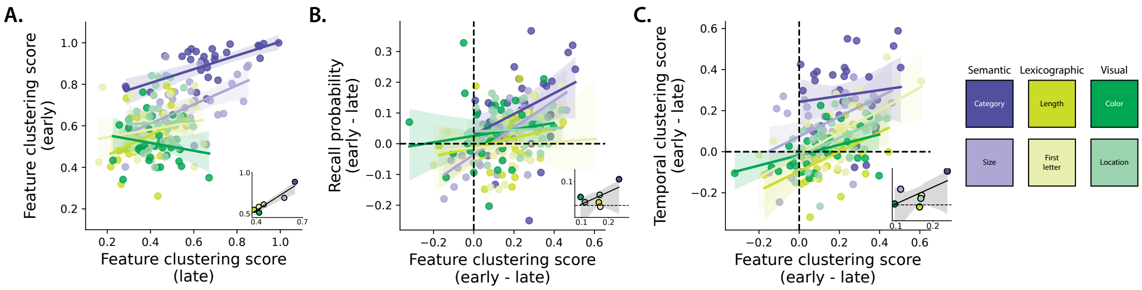

Figure 8: Feature clustering carryover effects. A. Feature clustering scores for order manipulation (early) versus randomly ordered (late) lists. **B.** Accuracy differences (on early versus late lists) versus feature clustering "carryover" (defined as the differences between the average clustering scores on early and late lists). **C.** Temporal clustering differences (on early versus late lists) versus feature clustering carryover. **All panels.** Each dot in the main scatterplots denotes the average scores for one participant. The colored regression lines are computed across participants. The inset displays condition-averaged results, where each dot reflects a single condition and the regression line is computed across experimental conditions. All error ribbons denote bootstrap-estimated 95% confidence intervals.

lists. We found that participants who exhibited larger carryover in feature clustering (i.e., continued to show strong feature clustering on late lists) for the semantic order manipulations (but not other manipulations) also tended to show a smaller decrease in recall on early versus late lists (Fig. 8B; overall: $r(179) = 0.307, p < 0.001, \mathrm{CI} = [0.148, 0.469]$; category: $r(28) = 0.350, p = 0.058, \mathrm{CI} = [0.050, 0.642]$; size: $r(28) = 0.708, p < 0.001, \mathrm{CI} = [0.472, 0.862]$; length: $r(28) = 0.205, p = 0.276, \mathrm{CI} = [-0.109, 0.492]$; first letter: $r(28) = 0.081, p = 0.672, \mathrm{CI} = [-0.433, 0.597]$; color: $r(29) = 0.155, p = 0.406, \mathrm{CI} = [-0.174, 0.541]$; location: $r(28) = 0.052, p = 0.787, \mathrm{CI} = [-0.307, 0.360]$; across conditions: $r(4) = 0.635, p = 0.176, \mathrm{CI} = [-0.924, 0.981]$. Participants who exhibited larger carryover in feature clustering also tended to show stronger temporal clustering on late lists (relative to early lists) for all but the category condition (Fig. 8C; overall: $r(179) = 0.426, p < 0.001, \mathrm{CI} = [0.285, 0.544]$; category: $r(28) = 0.110, p = 0.564, \mathrm{CI} = [-0.284, 0.442]$; all non-category conditions: all rs ≥ 0.406, all ps ≤ 0.023; across conditions: $r(4) = 0.649, p = 0.163, \mathrm{CI} = [-0.856, 0.988]$).

We suggest two potential interpretations of these findings. First, it is possible that some participants are more "malleable" or "adaptable" with respect to how they organize incoming information. When presented with list of items sorted along *any* feature dimension, they will simply adopt that feature as a dominant dimension for organizing those items and subsequent (randomly ordered) items. This flexibility in memory organization might afford such participants a memory advantage, explaining their strong recall performance. An alternative interpretation is that each participant comes into our study with a "preferred" way of organizing incoming information. If they happen to be assigned to an order manipulation condition that matches their preferences, then they will appear to be "sensitive" to the order manipulation and also exhibit a high degree of carryover in feature clustering from early to late lists. These participants might demonstrate strong recall performance not because of their inherently superior memory abilities, but rather because the specific condition they were assigned to happened to be especially easy for them, given their pre-experimental tendencies. To help distinguish between these interpretations, we designed an *adaptive* experimental condition (see *Adaptive condition*). The primary manipulation in the adaptive condition is that participants each experience three key types of lists. On *random* lists, words are ordered randomly (as in the feature-rich condition). On *stabilize* lists, the presentation order is adjusted to be maximally similar to the current estimate of the participant's memory fingerprint (see *Online "fingerprint" analysis*). Third, on *destabilize* lists, the presentation order is adjusted to be *minimally* similar to the current estimate of the participant's memory fingerprint (see *Ordering "stabilize" and "destabilize" lists by an estimated fingerprint*). The orders in which participants experienced each type of list were counterbalanced across participants to help reduce the influence of potential list-order effects. Because the presentation orders on stabilize and destabilize lists are adjusted to best match each participant's (potentially unique) memory fingerprint, the adaptive condition removes uncertainty about whether participants' assigned conditions

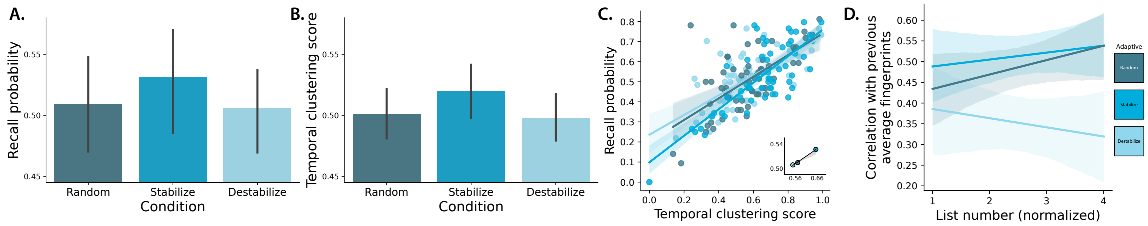

Figure 9: Adaptive free recall. A. Average probability of recall (taken across words, lists, and participants) for lists from each adaptive condition. **B.** Average temporal clustering scores for lists from each adaptive condition. **C.** Recall probability versus temporal clustering scores by participant (main panel; each participant contributes one dot per condition) and averaged within condition (inset; each dot represents a single condition). **D.** Per-list correlations between the current list's fingerprint and the average fingerprint computed from all previous lists. The normalized list numbers (x-axis) denote the number of lists of the same type that the participant had experienced at the time of the current list. All panels: Colors denote the sorting type (condition) for each list. Error bars and ribbons denote bootstrap-estimated 95% confidence intervals. For additional details about participants' behavior and performance during the adaptive conditions, see Figure S2.

might just "happen" to match their preferred ways of organizing their memories.

Participants' fingerprints on stabilize and random lists tended to become (numerically) slightly more similar to their average fingerprints computed from the previous lists they had experienced, and their fingerprints on destabilize lists tended to become numerically less similar (Fig. 9D). Overall, we found that participants tended to be better at remembering words on stabilize lists relative to words on both random ($t(59) = 1.740, p = 0.087, d = 0.095, \mathrm{CI} = [-0.187, 3.761]$) and destabilize ($t(59) = 1.714, p = 0.092, d = 0.114, \mathrm{CI} = [-0.351, 4.108]$) lists (Fig. 9A). Participants showed no reliable differences in their memory performance on destabilize versus random lists ($t(59) = -0.249, p = 0.804, d = -0.017, \mathrm{CI} = [-2.327, 1.578]$). Participants also exhibited stronger temporal clustering on stabilize lists, relative to random ($t(59) = 3.428, p = 0.001, d = 0.306, \mathrm{CI} = [1.635, 5.460]$) and destabilize ($t(59) = 4.174, p < 0.001, d = 0.374, \mathrm{CI} = [1.964, 6.968]$) lists (Fig. 9B). We found no reliable differences in temporal clustering for items on random versus destabilize lists ($t(59) = -0.880, p = 0.382, d = -0.081, \mathrm{CI} = [-3.165, 1.127]$).

As in the other experimental manipulations, participants in the adaptive condition exhibited substantial variability with respect to their overall memory performance and their clustering tendencies (Fig. 9C). We found that individual participants who exhibited strong temporal clustering scores also tended to recall more items. This held across subjects, aggregating across all list types ($r(178) = 0.701, p < 0.001, \mathrm{CI} = [0.590, 0.789]$), and for each list type individually (all rs ≥ 0.651, all ps < 0.001). Taken together, the results from the adaptive condition suggest that each participant comes into the experiment with their own unique memory organization tendencies, as characterized by their memory fingerprint. When participants study lists whose items come pre-sorted according to their unique preferences, they tend to remember more and show stronger temporal clustering.

We note that the multivariate aspect of the adaptive condition (i.e., sorting lists simultaneously along multiple feature dimensions) provides an important contrast with the order order manipulation conditions, where we sort lists along only a single feature dimension in each condition. We found that participants "naturally" clustered their recalls along multiple feature dimensions, even when the lists they studied were not sorted along those dimensions (as in the feature-rich condition). A caveat is that the *specific* feature dimensions participants tended to cluster along varied across participants. One way to quantify the multidimensional nature of participants' clustering tendencies is to sort each partipant's clustering scores (for each of the six feature dimensions, along with a seventh dimension to capture temporal clustering). We can then ask whether the distribution of clustering scores at each "rank" within the sorted set of scores for each participant has a mean that is reliably different from a chance value of 0.5. We carried out these tests for each set of ranked scores, and found that participants in the feature-rich condition reliably cluster their recalls along at least three dimensions, including temporal clustering (which was often ranked highest); Rank 1: $t(66) = 12.751, p < 0.001, d = 0.162, \mathrm{CI} = [8.702, 20.013]$; Rank 2: $t(66) = 8.196, p < 0.001, d = 0.162, \mathrm{CI} = [4.794, 12.978]$; Rank 3: $t(66) = 3.243, p =$

$0.002, d = 0.162, \text{CI} = [1.028, 7.051]$; Rank 4: $t(66) = -3.112, p = 0.003, d = 0.162, \text{CI} = [-5.282, -1.920]$; Rank 5: $t(66) = -7.154, p < 0.001, d = 0.162, \text{CI} = [-12.649, -5.568]$; Rank 6: $t(66) = -12.608, p < 0.001, d = 0.162, \text{CI} = [-22.114, -9.347]$; Rank 7: $t(66) = -18.397, p < 0.001, d = 0.162, \text{CI} = [-27.238, -14.073]$.

Discussion

We asked participants to study and freely recall word lists. The words on each list (and the total set of lists) were held constant across participants. For each word, we considered (and manipulated) two semantic features (category and size) that reflected aspects of the *meanings* of the words, along with two lexicographic features (word length and first letter), which reflected characteristics of the words' *letters*. These semantic and lexicographic features are intrinsic to each word. We also considered and manipulated two additional visual features (color and location) that affected the *appearance* of each studied item, but could be varied independently of the words' identities. Across different experimental conditions, we manipulated how the visual features varied across words (within each list), along with the orders of each list's words. Although the participants' task (verbally recalling as many words as possible, in any order, within one minute) remained constant across all of these conditions, and although the set of words they studied from each list remained constant, our manipulations substantially affected participants' memories. The impact of some of the manipulations also affected how participants remembered *future* lists that were sorted randomly.

Recap: visual feature manipulations

We found that participants in our feature-rich condition (where we varied words' appearances) recalled similar proportions of words to participants in a reduced condition (where appearance was held constant across words). However, varying the words' appearances led participants to exhibit much more temporal and feature-based clustering. This suggests that even seemingly irrelevant elements of our experiences can affect how we remember them.

When we held the within-list variability in participants' visual experiences fixed across lists (in the feature-rich and reduced conditions), they remembered more words from early lists than from late lists. For feature-rich lists, they also showed stronger clustering for early versus late lists. However, when we *varied* participants' visual experiences across lists (in the "reduced (early)" and "reduced (late)" conditions), these early versus late accuracy and clustering differences disappeared. Abruptly changing how incidental visual features varied across words seemed to act as a sort of "event boundary" that partially reset how participants processed and remembered post-boundary lists. Within-list clustering also increased in these manipulations, suggesting that the "within-event" words were being more tightly associated with each other.

When we held the visual features constant during early lists, but then varied words' appearances in later lists (i.e., the reduced (early) condition), participants' overall memory performance improved. However, this impact was directional: when we *removed* visual features from words in late lists that had been present in early lists (i.e., the reduced (late) condition), we saw no memory improvement.

Recap: order manipulations

When we (stochastically) sorted early lists along different feature dimensions, we found several impacts on participants' memories. Sorting early lists semantically (by word category) enhanced participants' memories for those lists, but the effects on performance of sorting along other feature dimensions were inconclusive. However, each order manipulation substantially affected how participants *organized* their memories of words from the ordered lists. When we sorted lists semantically, participants displayed stronger semantic clustering; when we sorted lists lexicographically, they displayed stronger lexicographic clustering; and when we sorted lists visually, they displayed stronger visual clustering. Clustering along the unmanipulated feature dimensions in each of these cases was unchanged.

The order manipulations we examined also appeared to induce, in some cases, a tendency to "clump" similar words within a list. This was most apparent on semantically ordered lists, where the probability of initiating recall with a given word seemed to follow groupings defined by feature change points.

We also examined the impact of early list order manipulations on memory for late lists. At the group level, we found little evidence for lingering "carryover" effects of these manipulations: participants in the order manipulation conditions showed similar memory performance and clustering on late lists to participants in the corresponding control (feature-rich) condition. At the level of individual participants, however, we found several meaningful patterns.

Participants who showed stronger feature clustering on early (order-manipulated) lists tended to better remember late (randomly ordered) lists. Participants who remembered early lists better also tended to show stronger feature clustering (along their condition's feature dimension) on late lists (even though the words on those late lists were presented in a random order). We also observed some (weaker) carryover effects of temporal clustering. Participants who showed stronger feature clustering (along their condition's feature dimension) on early lists tended to show stronger temporal clustering on late lists. And participants who showed stronger temporal clustering on early lists also tended to show stronger feature clustering on late lists. Essentially, these order manipulations appeared to affect each participant differently. Some participants were sensitive to our manipulations, and those participants' memory performance was impacted more strongly, both for the ordered lists and for future (random) lists. Other participants appeared relatively insensitive to our manipulations, and those participants showed little carryover effects on late lists.

These results at the individual participant level suggested to us that either (a) some participants were more sensitive to *any* order manipulation, or (b) some participants might be more (or less) sensitive to manipulations along *particular* (e.g., preferred) feature dimensions. To help distinguish between these possibilities, we designed an adaptive condition whereby we attempted to manipulate whether participants studied words in an order that either matched or mismatched our estimate of how they would cluster or organize the studied words in memory (i.e., their idiosyncratic memory fingerprint). We found that when we presented words in orders that were consistent with participants' memory fingerprints, they remembered more words overall and showed stronger temporal clustering. This comports well with the second possibility described above. Specifically, each participant seems to bring into the experiment their own idiosyncratic preferences and strategies for organizing the words in their memory. When we presented the words in an order consistent with each participant's idiosyncratic fingerprint, their memory performance improved. This might indicate that the participants were spending less cognitive effort "reorganizing" the incoming words on those lists, which freed up resources to devote to encoding processes instead.

Memory consequences of feature variability

Several prior studies have examined how varying the richness or experiences, or the extensive of encoding, can affect memory. Although specific details differ (Bonin et al., 2022), in general these studies have found that richer and more deeply or extensively encoded experiences are remembered better (Hargreaves et al., 2012; Madan, 2021; Meinhardt et al., 2020). Our findings help to elucidate an additional factor that may contribute to these phenomenon. For example, our finding that participants better remember "feature-rich" lists (where words' appearances are varied) than "reduced" lists (where words' appearances are held constant) only when those feature-rich lists are presented *after* reduced lists suggests that some factors that influence the richness or depth of encoding may be relative, rather than absolute. In other words, *increases* in richness (e.g., relative to a recency-weighted baseline) may be more important than the overall complexity or numbers of features.

Some prior studies have suggested that people can "cue" their memories using different "strategies" or "pathways" for searching for the target information. For example, modern accounts of free recall typically posit that memory search typically begins by matching the current state of mental context with the contexts associated with other items in memory (Kahana, 2020). Since context is the defining hallmark of episodic memory (Tulving, 1983), context-based search can be described as an "episodic" pathway to recall. When episodic cueing fails to elicit a match, participants may then search for items that are similar to the current mental context or mental state along other dimensions, such as semantic similarity (Davachi et al., 2003; Socher et al., 2009). These multiple pathways accounts of memory search also provide a potential explanation of why participants might have an easier time remembering richer stimuli (or experiences): richer stimuli and experiences might have more features that could be used to cue memory search. Our work suggests that there may be some additional factors at play with respect to the *dynamics* of these processes. In particular, we only observed memory benefits for "richer" stimuli when they were encountered after more "impoverished" stimuli (in the reduced (early) condition). This suggests that the pathways available to recall a given item may also depend on recent prior experiences.

We did *not* find any evidence that changing words' appearances *harmed* memory performance, e.g., by distracting them with irrelevant information (Lange, 2005; Marsh et al., 2012, 2015; Reinitz et al., 1992). Nor did we find any evidence that *changes* in the presence of potentially "distracting" features adversely affected memory. For example, when we increased or decreased the variability in words' appearances on late versus early lists (as in the reduced (early) and reduced (late) conditions), we found no evidence that this harmed participants' memories. One potential interpretation under the "multiple pathways to recall" framework is that the availability of multiple pathways to recall do not appear to specifically interfere with each other.

Context effects on memory performance and organization

In real-world experience, each moment's unique blend of contextual features (where we are, who we are with, what else we are thinking of at the time, what else we experience nearby in time, etc.) plays an important role in how we interpret, experience, and remember that moment, and how we relate it to our other experiences (e.g., for review see Manning, 2020). What are the analogues of real-world contexts in laboratory tasks like the free recall paradigm employed in our study? In general, modern formal accounts of free recall (Kahana, 2020) describe context as comprising a mix of (a) features pertaining to or associated with each item and (b) other items and thoughts experienced nearby in time, e.g., that might still be "lingering" in the participant's thoughts at the time they study the item. Item features can include semantic properties (i.e., features related to the item's meaning), lexicographic properties (i.e., features related to the item's letters), sensory properties (i.e., feature related to the item's appearance, sound, smell, etc.), emotional properties (i.e., features related to how meaningful the item is, whether the item evokes positive or negative feelings, etc.), utility-related properties (e.g., features that describe how an item might be used or incorporated into a particular task or situation), and more. Essentially any aspect of the participant's experience that can be characterized, measured, or otherwise described can be considered to influence the participant's mental context at the moment they experience that item. Temporally proximal features include aspects of the participant's internal or external experience that are *not* specifically occurring at the moment they encounter an item, but that nonetheless influence how they process the item. Thoughts related to percepts, goals, expectations, other experiences, and so on that might have been cued (directly or indirectly) by the participant's recent experiences prior to the current moment all fall into this category. Internally driven mental states, such as thinking about an experience unrelated to the experiment, also fall into this category.

Contextual features need not be intentionally or consciously perceived by the participant to affect memory, nor do they need to be relevant to the task instructions or the participant's goals. Incidental factors such as font color (Jones and Pyc, 2014), background color (Isarida and Isarida, 2007), inter-stimulus images (Chiu et al., 2021; Gershman et al., 2013; Manning et al., 2016), background sounds (Sahakyan and Smith, 2014; **?**), secondary tasks (Masicampto and Sahakyan, 2014; Oberauer and Lewandowsky, 2008; Polyn et al., 2009), and more can all impact how participants remember, and organize in memory, lists of studied items.

Consistent with this prior work, we found that participants were sensitive to task-irrelevant visual features. We also found that changing the dynamics of those task-irrelevant visual features (in the reduced (early) and reduced (late) conditions) *also* affected participants' memories. This suggests that it is not only the contextual features themselves that affect memory, but also the *dynamics* of context—i.e., how the contextual features associated with each item change over time.

Priming effects on memory performance and organization

When our ongoing experiences are ambiguous, we can draw on our past experiences, expectations, and other real, perceived, or inferred cues to help resolve these ambiguities. We may also be overtly or covertly "primed" to influence how we are likely to resolve ambiguities. For example, before listening to a story with several equally plausible interpretations, providing participants with "background" information beforehand can lead them towards one interpretation versus another (Yeshurun et al., 2017). More broadly, our conscious and unconscious biases and preferences can influence not only how we interpret high-level ambiguities, but even how we process low-level sensory information (Katabi et al., 2023).

In more simplified scenarios, like list-learning paradigms, the stimuli and tasks participants encounter before studying a given list can influence what and how they remember. For example, when participants are directed to suppress, disregard, or ignore "distracting" stimuli early on in an experiment, participants often tend to remember those stimuli less well when they are re-used as to-be-remembered targets later on in the experiment (Tipper, 1985). In general, participants' memories can be influenced by exposing them to a wide range of positive and negative priming factors before they encounter the to-be-remembered information (Balota et al., 1992; Clayton and Chattin, 1989; Donnelly, 1988; Flexser and Tulving, 1982; Gotts et al., 2012; Huang et al., 2004; Huber, 2008; Huber et al., 2001; McNamara, 1994; Neely, 1977; Rabinowitz, 1986; Tulving and Schacter, 1991; Watkins et al., 1992; Wiggs and Martin, 1998).

The order manipulation conditions in our experiment show that participants can also be primed to pick up on more subtle statistical structure in their experiences, like the dynamics of how the presentation orders of stimuli vary along particular feature dimensions. These order manipulations affected not only how participants remembered the manipulated lists, but also how they remembered *future* lists with different (randomized) temporal properties.

Free recall of blocked versus random categorized word lists

A large number of prior studies have compared participants' memories for categorized word lists that are presented in blocked versus random orders. In "blocked" lists, all of the words from a given semantic category (e.g., animals) are presented consecutively, whereas in "random" lists, the words from different categories are intermixed. Most of these studies report that participants tend to better remember blocked (versus random) lists (Bower et al., 1969; Cofer et al., 1966; D'Agostino, 1969; Dallett, 1964; Kintsch, 1970; Luek et al., 1971; Puff, 1974; Shapiro, 1970; **?**; **?**). Other studies suggest that these order effects may also be modulated by factors like list length and the numbers of exemplars in each category (e.g., Borges and Mangler, 1972).

Although we did not directly manipulate "blocking" in our order manipulation conditions, our sorting procedures in those conditions (see *Constructing feature-sorted lists*) have *indirect* effects on the lists' blockiness. For example, lists that are stochastically sorted by semantic category will tend to contain runs of several same-category words in succession. Consistent with the above work on blocked versus random categorized lists, we found that participants tended to better remember lists that were sorted semantically (Fig. 5B). However, this memory improvement did not appear to extend to the other order manipulation conditions we considered (e.g., to lexicographically or visually sorted lists). One possibility is that the memory benefits of blocked versus random lists are specific to semantic categories, and do not generalize to other feature dimensions. Another possibility is that the memory benefits are due to the presence of infrequent "jumps" between successive items (e.g., from different categories). Because the features we manipulated in the lexicographic and visual conditions were less categorical than the semantic features, feature values across words in those conditions tended to vary more gradually. Relatively stable features that are punctuated by infrequent large changes (e.g., as words transition from a same-category sequence to a new category) may also relate to perceived "event boundaries," which can have important consequences for memory (DuBrow and Davachi, 2013, 2016; DuBrow et al., 2017; Radvansky and Zacks, 2017).

Expectation, event boundaries, and situation models

Our findings that participants' current and future memory behaviors are sensitive to manipulations in which features change over time, and how features change across items and lists, suggest parallels with studies on how we form expectations and predictions, segment our continuous experiences into discrete events, and make sense of different scenarios and situations. Each of these real-world cognitive phenomena entail identifying statistical regularities in our experiences, and exploiting those regularities to gain insight, form inferences, organize or interpret memories, and so on. Our past experiences enable us to predict what is likely to happen in the future, given what happened "next" in our previous experiences that were similar to now (Barron et al., 2020; Brigard, 2012; Chow et al., 2016; Eichenbaum and Fortin, 2009; Gluck et al., 2002; Goldstein et al., 2021; Griffiths and Steyvers, 2003; Jones and Pashler, 2007; Kim et al., 2014; Manning, 2020; Tamir and Thornton, 2018; Xu et al., 2023).

When our expectations are violated, such as when our observations disagree with our predictions, we may perceive the "rules" or "situation" to have changed. *Event boundaries* denote abrupt changes in the state of our experience, for example, when we transition from one situation to another (Radvansky and Zacks, 2017; Zwaan and Radvansky, 1998). Crossing an event boundary can impair our memory for pre-boundary information and enhance our memory for post-boundary information (DuBrow and Davachi, 2013; Manning et al., 2016; Radvansky and Copeland, 2006; Sahakyan and Kelley, 2002). Event boundaries are also tightly associated with the notion of *situation models* and *schemas*—mental frameworks for organizing our understanding about the rules of how we and others are likely to behave, how events are likely to unfold over time, how different elements are likely to interact, and so on. For example, a situation model pertaining to a particular restaurant might set our expectations about what we are likely to experience when we visit that restaurant (e.g., what the building will look like, how it will smell when we enter, how crowded the restaurant is likely to be, the sounds we are likely to hear, etc.). Similarly, as mentioned in the *Introduction*, we might learn a schema describing how events are likely to unfold *across* any sit-down restaurant—e.g., open the door, wait to be seated, receive a menu, decide what to order, place the order, and so on. Situation models and schemas can help us to generalize across our experiences, and to generate expectations about how new experiences are likely to unfold. When those expectations are violated, we can perceive ourselves to have crossed into a new situation.

In our study, we found that abruptly changing the "rules" about how the visual appearances of words are determined, or about the orders in which words are presented, can lead participants to behave similarly to what one might expect upon crossing an event boundary. Adding variability in font color and presentation location for words on late lists, after those visual features had been held constant on early lists, led participants to remember more words on those later lists. One potential explanation is that participants perceive an "event boundary" to have occurred when they encounter the first "late" list. According to contextual change accounts of memory across event boundaries (e.g., Flores et al., 2017; Gold et al., 2017; Pettijohn et al., 2016; Sahakyan and Kelley, 2002), this could help to explain why participants in the reduced (early) condition exhibited better overall memory performance. Specifically, their memory for late list items could benefit from less interference from early list items, and the contextual features associated with late list items (after the "event boundary") might serve as more specific recall cues for those late items (relative to if the boundary had not occurred).

How do different types of clustering relate to each other, and to memory performance?

When the words on a studied list are presented in a random order, different types of clustering in participants' recalls often tend to be negatively correlated. For example, words that occur nearby on the list will not (on average) tend to be semantically related, and vice versa. Therefore a participant who shows a strong tendency to temporally cluster their recalls will tend to show weaker semantic clustering, and so on (Healey and Uitvlugt, 2019; Howard and Kahana, 2002b; Sederberg et al., 2010). Further, there is some evidence that temporal clustering is positively correlated with memory performance, whereas semantic clustering is negatively correlated with memory performance (Sederberg et al., 2010).

The notion of "multiple pathways to recall" discussed above (see *Memory consequences of feature variability*) suggests one potential explanation for these patterns. For example, temporal clustering has been proposed to reflect reliance on contextual cues in an "episodic" pathway to search memory, whereas semantic clustering reflects a relies on specific item features. These two pathways may "compete" with each other during recall (Socher et al., 2009). Meanwhile, extra-list intrusion errors (i.e., false "recalls" of items that were never encountered on the list) often tend to share semantic features with recently recalled items (Zaromb et al., 2006) and also often lead the participant to stop recalling additional items (Miller et al., 2012). Speculatively, over-reliance on semantic cues may lead to more intrusion errors, which in turn may lead to fewer recalls overall.

Our findings extend these prior results to consider lists that are *not* ordered randomly. Because ordering the words on a list along a particular feature dimension removes the "conflict" between temporal and feature clustering, the order manipulation conditions in our study represent an "edge case" whereby different pathways to recall are not necessarily in conflict with each other. For example, the same participants who exhibit strong feature clustering *also* show strong temporal clustering on ordered lists (Fig. 7E). This is presumably at least partly due to an inability to separate temporal and feature clustering on ordered lists (also see *Factoring out the effects of temporal clustering*). However, features that change gradually with time (i.e., presentation position) could also serve to strengthen the episodic (contextual) cues associated with each item. In other words, participants might essentially combine multiple noisy measures of change to form a more stable internal representation of temporal context.

Theoretical implications

Although most modern formal theories of episodic memory have been developed and tested to explain memory for list-learning tasks (Kahana, 2020), a number of recent studies suggest some substantial differences between memory for lists versus naturalistic stimuli (e.g., real-world experiences, narratives, films, etc.; Heusser et al., 2021; Lee et al., 2020; Manning, 2021; Nastase et al., 2020). One reason is that naturalistic stimuli are often much more engaging than the highly simplified list-learning tasks typically employed in the psychological laboratory, perhaps leading participants to pay more attention, exert more effort, and stay more consistently motivated to perform well (Nastase et al., 2020). Another reason is that the temporal unfoldings of events and occurrences in naturalistic stimuli tend to be much more meaningful than the temporal unfoldings of items on typical lists used in laboratory memory tasks. Real-world events exhibit important associations at a broad range of timescales. For example, an early detail in a detective story may prove to be a clue to solving the mystery later on. Further, what happens in one moment typically carries some predictive information about what came before or after (Xu et al., 2023). In contrast, the lists used in laboratory memory tasks are most often ordered randomly, by design, to *remove* meaningful temporal structure in the stimulus (Kahana, 2012).

On one hand, naturalistic stimuli provide a potential means of understanding how our memory systems function in the circumstances we most often encounter in our everyday lives. This implies that, to understand how memory works in the "real world," we should study memory for stimuli that reflect the relevant statistical structure of real-world experiences. On the other hand, naturalistic stimuli can be difficult to precisely characterize or model, making it difficult to distinguish whether specific behavioral trends follow from fundamental workings of our memory systems, from some aspect of the stimulus, or from idiosyncratic interactions or interference between participants' memory systems and the stimulus. This challenge implies that, to understand the fundamental nature of memory in its "pure" form, we should study memory for highly simplified stimuli that can provide relatively unbiased (compared with real-world experiences) measures of the relevant patterns and tendencies.

The experiment we report in this paper was designed to help bridge some of this gap between naturalistic tasks and more traditional list-learning tasks. We had people study word lists similar to those used in classic memory studies, but we also systematically varied the lists' "richness" (by adding or removing visual features) and temporal structure (through order manipulations that varied over time and across experimental conditions). We found that participants' memory behaviors were sensitive to these manipulations. Some of the manipulations led to changes that were common across people (e.g., more temporal clustering when words' appearances were varied, enhanced memory for lists following an "event boundary," more feature clustering on order-manipulated lists, etc.). Other manipulations led to changes that were idiosyncratic (especially carryover effects from order manipulations; e.g., participants who remembered more words on early order-manipulated lists tended to show stronger feature clustering for their condition's feature dimension on late randomly ordered lists, etc.). We also found that participants remembered more words from lists that were sorted to align with their idiosyncratic clustering preferences. Taken together, our results suggest that our memories are susceptible to external influences (i.e., to the statistical structure of ongoing experiences), but the effects of past experiences on future memory are largely idiosyncratic across people.

Potential applications

Every participant in our study encountered exactly the same words, split into exactly the same lists. But participants' memory performance, the orders in which they recalled the words, and the effects of early list manipulations on later lists all varied according to how we presented the to-be-remembered words.

Our findings raise a number of exciting questions. For example, how far might these manipulations be extended? In other words, might there be more sophisticated or clever feature or order manipulations that one could implement to have stronger impacts on memory? Are there limits to how much impact (on memory performance and/or organization) these sorts of manipulations can have? Are those limits universal across people, or are there individual differences (based on prior experiences, natural strategies, neuroanatomy, etc.) that impose person-specific limits on the potential impact of presentation-level manipulations on memory?

Our findings indicate that the ways word lists are presented affects how people remember them. To the extent that word list memory reflects memory processes that are relevant to real-world experiences, one could imagine potential real-world applications of our findings. For example, we found that participants remembered more words when the presentation order agreed with their memory fingerprints. If analogous fingerprints could be estimated for classroom content, perhaps they could be utilized manually by teachers, or even by automated content-presentation systems, to optimize how and what students remember.

Concluding remarks

Our work raises deep questions about the fundamental nature of human learning. What are the limits of our memory systems? How much does what we remember (and how we remember) depend on how we learn or experience the to-be-remembered content? We know that our expectations, strategies, situation models learned through prior experiences, and more collectively shape how our experiences are remembered. But those aspects of our memory are not fixed: when we are exposed to the same experience in a new way, it can change how we remember that experience, and also how we remember, process, or perceive *future* experiences.

Author contributions

Conceptualization: JRM and ACH. Methodology: JRM and ACH. Software: JRM, PCF, CEF, and ACH. Analysis: JRM, PCF, and ACH. Data collection: ECW, PCF, MRL, AMF, BJB, DR, and CEF. Data curation and management: ECW, PCF, MRL, and ACH. Writing (original draft): JRM. Writing (review and editing): ECW, PCF, MRL, AMF, BJB, DR, CEF, and ACH. Supervision: JRM and ACH. Project administration: ECW and PCF. Funding acquisition: JRM.

Author note

All of the data analyzed in this manuscript, along with all of the code for carrying out the analyses may be found at https://github.com/ContextLab/FRFR-analyses. Code for running the non-adaptive experimental conditions may be found at https://github.com/ContextLab/efficient-learning-code. Code for running the adaptive experimental condition may be found at https://github.com/ContextLab/adaptiveFR. We have also released an associated Python toolbox for analyzing free recall data, which may be found at https://cdl-quail.readthedocs.io/en/latest. Note that this study was not preregistered. Some of the ideas and data presented in this manuscript were also presented at the Annual Meeting of the Society for Neuroscience (2017) and the Context and Episodic Memory Symposium (2017).

Acknowledgements

We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of this study, and assistance with some of the data collection efforts from Rachel Chacko, Joseph Finkelstein, Sheherzad Mohydin, Lucy Owen, Gal Perlman, Jake Rost, Jessica Tin, Marisol Tracy, Peter Tran, and Kirsten Ziman. Our work was supported in part by NSF CAREER Award Number 2145172 to JRM. The content is solely the responsibility of the authors and does not necessarily represent the official views of our supporting organizations. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

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

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning and Motivation*, volume 2, pages 89–105. Academic Press, New York, NY.

Baddeley, A. D. (1968). Prior recall of newly learned items and the recency effect in free recall. *Canadian Journal of Psychology*, 22:157–163.

Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.

Balota, D. A., Black, S. R., and Cheney, M. (1992). Automatic and attentional priming in young and older adults: reevaluation of the two-process model. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):485–502.

Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a predictive coding account. *Progress in Neurobiology*, 192:101821–101834.

Bonin, P., Thiebaut, G., Bugaiska, A., and Méot, A. (2022). Mixed evidence for a richness-of-encoding account of animacy effects in memory from the generation-of-ideas paradigm. *Current Psychology*, 41:1653–1662.

Borges, M. A. and Mangler, G. (1972). Effect of within-category spacing on free recall. *Journal of Experimental Psychology*, 92(2):207–214.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 49:229–240.

Bousfield, W. A., Sedgewick, C. H., and Cohen, B. H. (1954). Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology*, 67:111–118.

Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11(2):177–220.

Bower, G. H., Lesgold, A. M., and Tieman, D. (1969). Grouping operations in free recall. *Journal of Verbal Learning and Verbal Behavior*, 8(4):481–493.

Brigard, F. D. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*, 3(420):1–3.

Chiu, Y.-C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., and Sahakyan, L. (2021). Separation of item and context in item-method directed forgetting. *NeuroImage*, 235:117983.

Chow, W.-Y., Momma, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.

Clayton, K. and Chattin, D. (1989). Spatial and semantic priming effects in tests of spatial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3):495–506.

Clewett, D., DuBrow, S., and Davachi, L. (2019). Transcending time in the brain: how event memories are constructed from experience. *Hippocampus*, 29(3):162–183.

Cofer, C. N., Bruce, D. R., and Reicher, G. M. (1966). Clustering in free recall as a function of certain methodological variations. *Journal of Experimental Psychology: General*, 71:858–866.

D'Agostino, P. R. (1969). The blocked-random effect in recall and recognition. *Journal of Verbal Learning and Verbal Behavior*, 8:815–820.

Dallett, K. M. (1964). Number of categories and category information in free recall. *Journal of Experimental Psychology*, 68:1–12.

Darley, C. F. and Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology: General*, 91:66–73.

Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences, USA*, 100(4):2157–2162.

Donnelly, R. E. (1988). Priming effects in successive episodic tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:256–265.

Drewnowski, A. and Murdock, B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory*, 6:319–332.

DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.

DuBrow, S. and Davachi, L. (2016). Temporal binding within and across events. *Neurobiology of Learning and Memory*, 134:107–114.

DuBrow, S., Rouhani, N., Niv, Y., and Norman, K. A. (2017). Does mental context drift or shift? *Current Opinion in Behavioral Sciences*, 17:141–146.

Eichenbaum, H. and Fortin, N. J. (2009). The neurobiology of memory based predictions. *Philosophical Transactions of the Royal Society of London Series B*, 364(1521):1183–1191.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, 62:145–154.

Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science*, 22(2):243–252.

Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall: a challenge for unitary models of recency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36:324–347.

Farrell, S. (2014). Correcting the correction of conditional recency slopes. *Psychonomic Bulletin and Review*, 21:1174–1179.

Fitzpatrick, P. C., Heusser, A. C., and Manning, J. R. (2023). Text embedding models yield high-resolution insights into conceptual knowledge from short multiple-choice quizzes. *PsyArXiv*, page doi.org/10.31234/osf.io/dh3q2.

Flexser, A. J. and Tulving, E. (1982). Priming and recognition failure. *Journal of Verbal Learning and Verbal Behavior*, 21:237–248.

Flores, S., Bailey, H. R., Eisenberg, M. L., and Zacks, J. M. (2017). Event segmentation improves event memory up to one month later. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(8):1183.

Gershman, S. J., Schapiro, A. C., Hupbach, A., and Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *The Journal of Neuroscience*, 33(20):8590–8595.

Glenberg, A. M., Bradley, M. M., Kraus, T. A., and Renzaglia, G. J. (1983). Studies of the long-term recency effect: support for a contextually guided retrieval theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12:413–418.

Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the "weather prediction" task? individual variability in strategies for probabilistic category learning. *Learning and Memory*, 9:408–418.

Gold, D. A., Zacks, J. M., and Flores, S. (2017). Effects of cues to event segmentation on subsequent memory. *Cognitive Research: Principles and Implications*, 2(1):1.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. (2021). Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*, page doi.org/10.1101/2020.12.02.403477.

Gotts, S. J., Chow, C. C., and Martin, A. (2012). Repetition priming and repetition suppression: A case for enhanced efficiency through neural synchronization. *Cognitive Neuroscience*, 3(3-4):227–237.

Griffiths, T. L. and Steyvers, M. (2003). Prediction and semantic association. *Advances in Neural Information Processing Systems*, 15.

Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml, M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages 2338–2342.

Hargreaves, I. S., Pexman, P. M., Johnson, J. C., and Zdrazilova, L. (2012). Richer concepts are better remembered: number of features effects in free recall. *Frontiers in Human Neuroscience*, 6:doi.org/10.3389/fnhum.2012.00073.

Healey, M. K. and Uitvlugt, M. G. (2019). The role of control processes in temporal and semantic contiguity. *Memory and Cognition*, 47:719–737.

Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a Python toolbox for analyzing and plotting free recall data. *Journal of Open Source Software*, 10.21105/joss.00424.

Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral and neural signatures of transforming experiences into memories. *Nature Human Behavior*, 5:905–919.

Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). HyperTools: a Python toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning Research*, 18(152):1–6.

Hogan, R. M. (1975). Interitem encoding and directed search in free recall. *Memory and Cognition*, 3:197–209.

Howard, M. W. and Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:923–941.

Howard, M. W. and Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46:269–299.

Howard, M. W. and Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, 46:85–98.

Huang, L., Holcombe, A. O., and Pashler, H. (2004). Repetition priming in visual search: episodic retrieval, not feature priming. *Memory and Cognition*, 32:12–20.

Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental Psychology: General*, 137(2):324–347.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., and Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, 108(1):149–182.

Isarida, T. and Isarida, T. K. (2007). Environmental context effects of background color in free recall. *Memory and Cognition*, 35(7):1620–1629.

Jenkins, J. J. and Russell, W. A. (1952). Associative clustering during recall. *Journal of Abnormal and Social Psychology*, 47:818–821.

Jones, A. C. and Pyc, M. A. (2014). The production effect: costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):300–305.

Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.

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

Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York, NY.

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, 71:107–138.

Kahana, M. J., Howard, M. W., and Polyn, S. M. (2008). Associative processes in episodic memory. In Roediger III, H. L., editor, *Cognitive Psychology of Memory*, pages 476–490. Elsevier, Oxford, UK.

Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., and Yeshurun, Y. (2023). Deeper than you think: partisanship-dependent brain responses in early sensory and motor brain regions. *The Journal of Neuroscience*, pages doi.org/10.1523/JNEUROSCI.0895–22.2022.

Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences, USA*, In press.

Kimball, D. R., Smith, T. A., and Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review*, 114(4):954–993.

Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.

Lange, E. B. (2005). Disruption of attention by irrelevant stimuli in serial recall. *Journal of Memory and Language*, 43(4):513–531.

Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases of human memory. *Current Opinion in Behavioral Sciences*, 32:111–119.

Lohnas, L. J., Polyn, S. M., and Kahana, M. J. (2010). Modeling intralist and interlist effects in free recall. In *Psychonomic Society*, Saint Louis, MO.

Luek, S. P., Mclaughlin, J. P., and Cicala, G. A. (1971). Effects of blocking of input and blocking of retrieval cues on free recall learning. *Journal of Experimental Psychology*, 91(1):159–161.

Madan, C. R. (2021). Exploring word memorability: how well do different word properties explain item free-recall probability? *Psychonomic Bulletin and Review*, 28:583–595.

Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook of Human Memory*. Oxford University Press.

Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave" function? *Psychological Review*, 128(4):711–725.

Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. A. (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic Bulletin and Review*, 23(5):1534–1542.

Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free recall. *Memory*, 20(5):511–517.

Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.

Manning, J. R., Notaro, G. M., Chen, E., and Fitzpatrick, P. C. (2022). Fitness tracking reveals task-specific associations between memory, mental health, and physical activity. *Scientific Reports*, 12(13822):doi.org/10.1038/s41598–022–17781–0.

Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, 108(31):12893–12897.

Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., and Kahana, M. J. (2012). Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clustering during memory search. *The Journal of Neuroscience*, 32(26):8871–8878.

Marsh, J. E., Beaman, C. P., Hughes, R. W., and Jones, D. M. (2012). Inhibitory control in memory: evidence for negative priming in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5):1377–1388.

Marsh, J. E., Sörqvist, P., Hodgetts, H. M., Beaman, C. P., and Jones, D. M. (2015). Distraction control processes in free recall: benefits and costs to performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(1):118–133.

Masicampto, E. J. and Sahakyan, L. (2014). Imagining another context during encoding offsets context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6):1772–1777.

Masís-Obando, R., Norman, K. A., and Baldassano, C. (2022). Scheme representations in distinct brain networks support narrative memory during encoding and retrieval. *eLife*, 11:e70445.

McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:507–520.

Meinhardt, M. J., Bell, R., Buchner, A., and Röer, J. P. (2020). Adaptive memory: is the animacy effect on memory due to richness of encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(3):416–426.

Miller, J. F., Kahana, M. J., and Weidemann, C. T. (2012). Recall termination in free recall. *Memory and Cognition*, 40(4):540–550.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behavior*, 1:680–692.

Moran, R. and Goshen-Gottstein, Y. (2014). The conditional-recency dissociation is confounded with nominal recency: should unitary models of memory still be devaluated? *Psychonomic Bulletin and Review*, 21:332–343.

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

Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106(3):226–254.

Oberauer, K. and Lewandowsky, S. (2008). Forgetting in immediate serial recall: decay, temporal distinctiveness, or interference? *Psychological Review*, 115(3):544–576.

Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., and Radvansky, G. A. (2016). Event boundaries and memory improvement. *Cognition*, 148:136–144.

Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of context. *Trends in Cognitive Sciences*, 12:24–30.

Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). Task context and organization in free recall. *Neuropsychologia*, 47:2158–2163.

Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, 17:132–138.

Puff, C. R. (1974). A consolidated theoretical view of stimulus-list organization effects in free recall. *Psychological Reports*, 34:275–288.

Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In Bower, G. H., editor, *The Psychology of Learning and Motivation: Advances in Research and Theory*, volume 14, pages 207–262. Academic Press, New York, NY.

Rabinowitz, J. C. (1986). Priming in episodic memory. *Journal of Gerontology*, 41:204–213.

Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting: situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.

Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, 17:133–140.

Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature Reviews Neuroscience*, 13:713–726.

Reinitz, M. T., Lammers, W. J., and Cochran, B. P. (1992). Memory-conjunction errors: miscombination of stored stimulus features can produce illusions of memory. *Memory and Cognition*, 20:1–11.

Rissman, J., Eliassen, J. C., and Blumstein, S. E. (2003). An event-related fMRI investigation of implicit semantic priming. *Journal of Cognitive Neuroscience*, 15(8):1160–1175.

Romney, A. K., Brewer, D. D., and Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, 4:28–34.

Sahakyan, L. and Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6):1064–1072.

Sahakyan, L. and Smith, J. R. (2014). A long time ago, in a context far, far away: Retrospective time estimates and internal context change. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):86–93.

Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclopedic Reference*, 3:501–506.

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

Sederberg, P. B., Miller, J. F., Howard, W. H., and Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory and Cognition*, 38(6):689–699.

Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, 24:134–193.

Shapiro, S. I. (1970). Isolation effects, free recall, and organization. *Journal of Psychology*, 24:178–183.

Sirotin, Y. B., Kimball, D. R., and Kahana, M. J. (2005). Going beyond a single list: modeling the effects of prior experience on episodic free recall. *Psychonomic Bulletin and Review*, 12(5):787–805.

Slamecka, N. J. and Barlow, W. (1979). The role of semantic and surface features in word repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 18:617–627.

Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: a review and meta-analysis. *Psychonomic Bulletin and Review*, 8(2):203–220.

Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009). A Bayesian analysis of dynamics in free recall. *Advances in Neural Information Processing Systems*, 22.

Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., and Zacks, J. M. (2011). Changes in events alter how people remember recent information. *Journal of Cognitive Neuroscience*, 23(5):1052–1064.

Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, 138(2):236–257.

Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, 22(3):201–212.

Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37:571–590.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., and Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, 316(5821):76–82.

Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press, New York, NY.

Tulving, E. and Schacter, D. L. (1991). Priming and human memory systems. *Science*, 247:301–305.

Watkins, P. C., Mathews, A., Williamson, D. A., and Fuller, R. D. (1992). Mood-congruent memory in depression: emotional priming or elaboration? *Journal of Abnormal Psychology*, 101(3):581–586.

Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal of Psychology*, 35:396–401.

Whitely, P. L. (1927). The dependence of learning and recall upon prior intellectual activities. *Journal of Experimental Psychology: General*, 10:489–508.

Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, 8(2):227–233.

Xu, X., Zhu, Z., and Manning, J. R. (2023). The psychological arrow of time drives temporal asymmetries in retrodicting versus predicting narrative events. *PsyArXiv*, page doi.org/10.31234/osf.io/yp2qu.

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., and Hasson, U. (2017). Same story, different story: the neural representation of interpretive frameworks. *Psychological Science*, 28(3):307–319.

Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., and Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792–804.

Zhang, Q., Griffiths, T. L., and Norman, K. A. (2023). Optimal policies for free recall. *Psychological Review*, 130(4):1104–1125.

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, 50:2597–2605.

Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation models in narrative comprehension: an event-indexing model. *Psychological Science*, 6(5):292–297.

Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.