Carryover effects in free recall reveal how prior experiences influence memories of new experiences

Jeremy R. Manning[1,*], Kirsten Ziman[1,2], Emily Whitaker[1],
Paxton C. Fitzpatrick[1], Madeline R. Lee[1], Allison M Frantz[1],
Bryan J. Bollinger[1], Campbell E. Field[1], and Andrew C. Heusser[1,3]

[1]Dartmouth College

[2]Princeton University

[3]Akili Interactive

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

Abstract

We perceive, interpet, and remember ongoing experiences through the lens of our prior experiences. Inferring that we are 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 of 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 the orders in which the words were studied. We found that these order manipulations affected not only how the participants recalled the ordered lists, but also how they recalled later randomly ordered lists. Our work shows how structure in our ongoing experiences can exert influence on how we remember unrelated subsequent experiences.

Introduction

Experience is subjective: different people who encounter identical physical experiences can take away very different meanings and memories. One reason is that our subjective experiences in the moment are shaped in part 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; Ranganath and Ritchey, 2012) or *schemas* (Baldassano et al., 2018; Masís-Obando et al., 2022) 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 potentail ambiguities in their ongoing experiences, including ambiguous movies and narratives (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 studied using more traditional paradigms like free recall of random word lists (Kahana, 2012). In free recall, 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.

Nevertheless, there are *some* commonalities between memory for word lists and memory for real-world experiences.

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 the existence of a special representation, called *context*, 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.

Over the past half-century, context-based models have enjoyed 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, 2002; 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 study list), as well as semantic and temporal clustering effects (**?**). 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 the tendency for people to successively recall items that occupied neighboring positions in the study list. For example, if a list contained the sub-sequence "ABSENCE HOLLOW PUPIL" and the participant recalls the word "HOLLOW", it is far more likely that the next response will be either "PUPIL" or "ABSENCE" than some other list item (Kahana, 1996). In addition, there is a strong forward bias in the contiguity effect: subjects make forward transitions (i.e., "HOLLOW" followed by "PUPIL") about twice as often as they make backward transitions, despite an overall tendency to begin recall at the end of the list. 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. As captured by 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, font size, font color, location on the screen, size of the object the word represents, 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 recall in succession words that share stimulus features.

A key mystery is whether 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, situation models and clustering effects both reflect statistical regularities in ongoing experience. 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., 2022). On the other hand, the rich structure of real-world experiences and other naturalistic stimuli that enable people to form deep and meaningful situation models and schemas have no obvious analog 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 change across different real-world situations. We asked participants to study and freely recall a series of word lists (Fig. 1). Across the different conditions in the experiment, we varied the lists' presentation orders in different ways across lists. 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). In our main manipulation conditions, we asked participants to study and recall eight lists whose items were sorted by a target feature (e.g., word category). Next, we asked them to study and recall an additional eight lists whose items had the same features, but that were sorted in a random temporal order. We were interested in how these order manipulations affected participants' recall behaviors on early (sorted) lists, as well as how order manipulations on early lists affected recall behaviors on later (unsorted) lists. We used a series of 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. Finally, in an *adaptive* experimental condition we used participants' recall behaviors on early lists to manipulate, in real-time, the presentation orders of subsequent lists. In this adaptive condition, we sought to identify potential commonalities within and across participants in how people organized their memories and how those organizational tendancies affect overall performance.

Materials and methods

Participants

We enrolled a total of 491 Dartmouth undergraduate students across 11 experimental conditions. The conditions included two primary controls (feature rich, reduced), two secondary controls (reduced (early), reduced (late)), six order manipulation conditions (category, size, length, first letter, color, and location), and a final adaptive condition. Each of these conditions are described in the *Experimental design* subsection below.

Participants received course credit for enrolling in our study. We asked each participant to fill out a demographic survey that included information about their self-reported age, gender, ethnicity, race, education, vision, reading impairments, medications or 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 report some or all of their requested demographic information.

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 other conditions we set a target enrollment of at least 30 participants. Because our data collection efforts were coordinated 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.

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; standard deviation: 1.356). A total of 318 participants reported their gender as female, 170 as male, and 2 participants declined to report their gender. A total of 442 participants reported their ethnicity as "not Hispanic or Latino," 39 as "Hispanic or Latino," and 9 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 particiapnts), Native Hawaiian or Other Pacific Islander (4 participants), Mixed race (3 participants), Middle Eastern (1 participant), and Arab (1 participant). A total of 5 participants declined to report their race. We note that several participants reported more than one of racial category. Participants reported their highest degrees achieved as "Some college" (359 participants), "High school graduate" (117 participants), "College graduate" (7 participants), "Some high school" (5 participants), "Doctorate" (1 participant), and "Master's degree" (1 participant). A total of 482 participants reported no reading impairments, and 8 reported mild reading impairments such as mild dyslexia. A total of 489 participants reported having normal color vision and 1 participant reported that they were color blind. A total of 482 participants reported taking no prescription medications and having no recent injuries; 4 participants reported having ADHD, 1 reported having dyslexia, 1 reported having allergies, 1 reported a recently torn ACL/MCL, and 1 reported a concussion from several months prior. The participants reported consuming 0 – 3 cups of coffee prior to 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 1 participant who reported abnormal color vision, as well as 39 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 (late) (41 participants), reduced (early), (42 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.

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, referrent 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 (text color, presentation location). Each list contains four words from each of four different semantic categories and two object sizes; all other stimulus features are randomized. 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 evaluate participants' memory fingerprints in rich detail.

Stimuli

Stimuli in our paradigm were 256 English words selected in a previous study (Ziman et al., 2018). The words all 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 the word referred to. Words were labeled as "small" if the corresponding object was likely able to "fit in a standard shoebox" or "large" if the object was larger than a shoebox. Semantic categories varied in how many object sizes they reflected (mean number of different sizes per category: 1.33; standard deviation: 0.49). The numbers of words in each semantic category also varied from 12 – 28 (mean number of words per category: 17.07; standard deviation number of words: 4.65). We also identified lexicographic features for each word, including the words' first letters and lengths (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 contained words from exactly 4 unique categories, each with exactly 4 examplars from each category. Second, we required that (across all words on the list) at least one instance of both object sizes were represented. On average, each category was represented in 4.27 lists (standard deviation: 1.16 lists). Aside from these

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 the first lists 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.

two constraints, we assigned each word to a unique list. After random assignment, each list contained words with an average of 11.13 unique starting letters (standard deviation: 1.15 letters) and an average word 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 some conditions, on some lists, we also randomly varied two additional visual features to each word: the presentation font color, and the word's onscreen location. These attributes were assigned independently for word (and for every participant) at the times the words were displayed onscreen. 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, words were all presented in black at the center of the experimental computer's display.

To assign a random font color to each word, we selected three integers uniformly and at random between 0 and 255, 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 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, using this terminology 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 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 four "control" conditions to evaluate and explore participants' baseline behaviors. We also used performance on these control conditions to help interpret performance in other "manipulation" conditions. Two control conditions served as "anchorpoints." In the first anchorpoint 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 anchorpoint 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.

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 earlylists 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, some features were correlated– for example, some semantic categories of words referred to objects that tended to be a particular size, which means that category and size are not fully independent. 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 each pair of features. This function varied according to the type of features. 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 are 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 for that feature. 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 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" takes on multiple (positive) real values. We defined the "color" and "location" distances between two words as the Euclidean distances between their (r, g, b) color or (x, y) location vectors, respectively. Therefore the color and location distance measures always take on positive real values (upper bounded at 441.67 for color, or 27 in 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. First, we choose a word uniformly at random from the set of candidates. Next, 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:

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

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 computed a set of normalized similarity values by dividing the similarities by their sum:

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

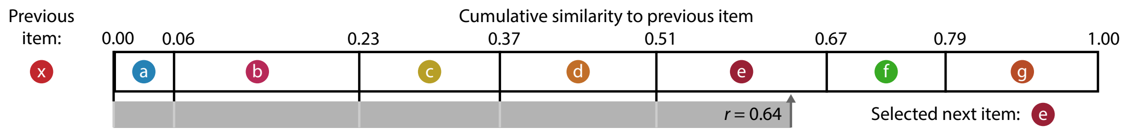

Figure 2: Generating stochastic feature-sorted lists. For a given feature dimension (e.g., color), we compute the similarity (Eqn. 1) between the feature value(s) of the previous item, x, and all yet-to-be-presented items (a – g). Next, we normalize these similarity scores so that they sum to one. We lay in sequence a set of "sticks," one for each candidate item, whose lengths are equal to these normalized similarity scores. Note that the combined lengths of these sticks is one. To select the next to-be-presented item, we draw a random number, r, from the uniform distibution 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.

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

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 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 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.

Adaptive condition

We designed the *adaptive* experimental condition to study the effect on memory for information that matched (or mismatched) the ways participants "naturally" organized their memories of the lists they studied. Like the other conditions, all participants in the adaptive condition studied a total of 16 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 and features 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," defined below. At a high level, a participant's memory fingerprint describes how they prioritize different semantic, lexicographic, and/or visual features when they organize their memories.

Next, participants studied a sequence of 12 lists in three batches of 4 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 computing 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 base 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 those feature values from the first word the participant recalled on the 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 averaged 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 repeat 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 episode events in the exact order they occurred (or in exact reverse order), this would yield a score of 1. If a participant recalled the events in random order, this would yield an expected score of 0.5. For each recall-event transition (and separately for each participant), we sorted all not-yet-recalled events according to their absolute lag (that is, distance away in the episode). We then computed the percentile rank of the next event 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 *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. And 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 orders and numbers of correct recalls a participant makes, the number of itervening 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. The 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-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 compute a "null" distribution of n additional clustering scores after randomly shuffling the recall order (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 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.

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, this memory fingerprint describes the participant's tendancies to order (and, presumably, organize in memory) the studied words along each dimension. To obtain stable estimates of these fingerprints for each participant, we averaged 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 participants' *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 alorithms to the just collected 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 averaged across lists, the already-computed memory fingerprints for earliar lists could be cached and loaded as needed in future computations. This meant that our computations pertaining to updating our estimates of a participant's memory fingerprint only needed to consider data from the most recent list. Second, each element of the null distributions of uncorrected fingerprint scores (see *Permutation-corrected feature clustering scores*) could be estimated independently from the others. This enabled us to make use of the testing computers' multi-core CPU archetectuers by elements of the null distributions in batches of eight (i.e., the number of CPU cores on each testing computer). Taken together, we were able to compress the fingerprint computations into just a few seconds of computing time. The combined processing time for the speech-to-text algorithm and fingerprint computations easily fit within the inter-list intervals, where participants typically paused before moving on to 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 prospective presentation orders. We sought to select the specific order that most (or least) 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.) 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 stabilized 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 (Heusser et al., 2018), we use 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 fingerprints (across all lists, participants, and experimental conditions) to create a large matrix with number-of-lists $\times$ number-of-participants rows and seven columns (one for each word feature dimension's clustering scores, 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, and S8B) we computed an additional set of two-dimensional embeddings for participants' *average* fingerprints (i.e., across lists within a given group of lists– early or late). For those visualizations we averaged each participant's rows (for the given 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 *participant* and *list group*, 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.

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. To carry out this analysis, we initialized (for each participant) a number-of-lists (16) by number-of-words-per-list (16) matrix of zeros. Then, for each list, we found the index of the word that was recalled first, and we 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 probabilility of n^{th} reacall 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 paritcipant's curve(s).

Lag-conditional response probability curve

The lag-conditional probability (lag-CRP) curve (Kahana, 1996) reflects the probability of recalling a given item after the just-recalled item, as a function of their 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 3 items before the previously recalled item. For each recall transition (following the first recall), we computed the lag between the just-recalled word's presentation position and the next-recalled word's presentation position. We 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 successive repetitions (e.g., recalling the same word twice in a row). 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.

Serial position curve

Serial position curves (Murdock, 1962) reflect the proportion of participants who remember each item as a function of the item's serial position during encoding. For each participant, we initialized a number-of-lists (16) by number-of-words-per-list (16) matrix of zeros. 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 one or zero). 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 occuring 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).

Results

We sought to manipulate two aspects of how participants memorized sequences of word lists. 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 refferent) and of the lexicographic properties of the word (e.g., word length, first letter). We wondered whether this additional word-independent information might facilitate recall (e.g., by providing new potential ways of organizing or retrieving memories of the studied words) or impair recall (e.g., by distracting participants). Second, our primary experimental manipulations entailed manipulating the orders in which words were studied (and how those orderings changed over time). We wondered whether presenting the same list of words in different orders (e.g., sorted along one feature dimension versus another) might serve to influence how participants organized their memories of the words. We also wondered whether some order manipulations might be temporally "sticky" by influencing how *future* lists were remembered.

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 for feature rich versus reduced lists ($t(126) = -0.290, p = 0.772$). However, participants in the feature rich condition clustered their recalls substantially more along every dimension we examined (temporal clustering: $t(126) = 10.624, p < 0.001$; category clustering: $t(126) = 10.077, p < 0.001$; size clustering: $t(126) = 11.829, p < 0.001$; word length clustering: $t(126) = 10.639, p < 0.001$; first letter clustering: $t(126) = 7.775, p = 0.000$; 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).

We also wondered whether adding these irrelevant 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: $t(66) = 4.553, p < 0.001$; reduced: $t(60) = 2.434, p = 0.018$). Participants in the feature rich (but not reduced) conditions exhibited more temporal clustering on early versus late lists (feature rich: $t(66) = 2.318, p = 0.024$; reduced: $t(60) = 0.929, p = 0.357$). And participants in both conditions exhibited more semantic (category and size) clustering on early versus late lists (feature rich, category: $t(66) = 3.805, p < 0.001$; feature rich, size: $t(66) = 2.190, p = 0.032$; reduced, category: $t(60) = 2.856, p = 0.006$; reduced, size: $t(60) = 2.947, p = 0.005$). Participants in the reduced (but not feature rich) conditions exhibited more lexicographic clustering on early versus late lists (feature rich, word length: $t(66) = 0.161, p = 0.872$; feature rich, first letter: $t(66) = 0.410, p = 0.683$; reduced, word length: $t(60) = 3.528, p = 0.001$; reduced, first letter: $t(60) = 2.275, p = 0.026$). Taken together, these comparisons suggest that even when the presence or absence of irrelevant 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 irrelevant visual features constant on early lists, but allowed them to vary randomly on late lists. In a *reduced (late)* condition, we allowed the irrelevant visual features to vary randomly on early lists, but held them constant on late lists. Given our above findings that (a) participants tended to remember more words and 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): $t(41) = 1.499, p = 0.141$; reduced (late): $t(40) = 1.462, p = 0.152$), temporal clustering (reduced (early): $t(41) = 0.998, p = 0.324$; reduced (late):

$t(40) = 1.099, p = 0.278$), nor feature based clustering (reduced (early), category: $t(41) = 0.753, p = 0.456$; reduced (early), size: $t(41) = 0.721, p = 0.475$; reduced (early), length: $t(41) = 0.493, p = 0.625$; reduced (early), first letter: $t(41) = 0.780, p = 0.440$; reduced (late), category: $t(40) = -0.086, p = 0.932$; reduced (late), size: $t(40) = 0.746, p = 0.460$; reduced (late), length: $t(40) = 1.476, p = 0.148$; reduced (late), first letter: $t(40) = 0.966, p = 0.340$). We hypothesized that adding or removing the irrelevant features was acting as a sort of "event boundary" between early and late lists. 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; Manning et al., 2016).

We found that *adding* irrelevant 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): $t(107) = -2.230, p = 0.028$; reduced versus reduced (early): $t(101) = -2.045, p = 0.043$; 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): $t(106) = -0.638, p = 0.525$; reduced versus reduced (late): $t(100) = -0.407, p = 0.685$). These comparisons suggest that recall accuracy has a directional component (i.e., 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 irrelevant visual features to *any* lists (comparisons of clustering on feature rich and reduced lists are reported above; temporal clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions: ts ≤ -9.780, ps < 0.001; feature based clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions: ts ≤ -5.443, 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: $t\mathrm{s} \geq -1.434$, $p\mathrm{s} \geq 0.154$; feature based clustering in feature rich versus reduced (early) and feature rich versus reduced (late) conditions: $t\mathrm{s} \geq -1.359$, $p\mathrm{s} > 0.177$).

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 while holding the features themselves constant. We hypothesized that changing the order 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 visual features were added or removed from a subset of the lists participants studied).

Across six order manipulation conditions, we sorted early lists by each 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-ordered

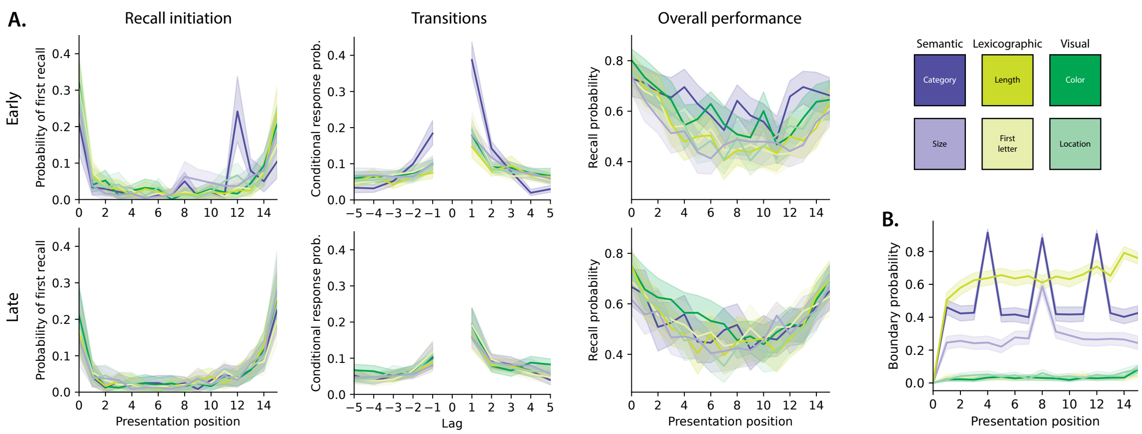

Figure 3: Recall dynamics in feature rich free recall (order manipulation conditions). A. Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling each word 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 (control) and adaptive conditions. **B.** Proportion of event boundaries (see *Identifying event boundaries*) for each condition's feature of focus, plotted as a function of presentation position.

condition showed an increase in memory performance on early lists (accuracy, relative to early feature rich lists; $t(95) = 3.034, p = 0.003$). Participants in the color-ordered condition also showed a trending increase in memory performance on early lists (again, relative to early feature rich lists: $t(96) = 1.850, p = 0.067$). Participants' performance on early lists in all of the other order manipulation conditions was indistinguishable from performnace on the early feature rich lists ($||t||\text{s} < 1.013, p\text{s} > 0.314$). Participants in both of the semanticly ordered conditions exhibited stronger temporal clustering on early lists (versus early feature rich lists; category: $t(95) = 8.508, p < 0.001$; size: $t(95) = 2.429, p = 0.017$). Participants in the length-ordered condition tended to exhibit *less* temporal clustering on early lists relative to early feature rich lists ($t(95) = -1.666, p = 0.099$), whereas participants in the first letter-ordered condition exhibited stronger temporal clustering on early lists ($t(95) = 2.587, p = 0.011$). Participants in the visually ordered conditions exhibited more similar performance on early lists, relative to early feature rich lists (color: $t(96) = -1.064, p = 0.290$; we found a trending enhancement for participants in the location-ordered condition: $t(95) = 1.682, p = 0.096$). We also compared feature-based clustering on early lists across the order manipulation and feature rich conditions. Since results were similar across both semantic conditoins (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: $t(125) = 2.524, p = 0.013$; size:$t(125) = 3.510, p = 0.001$), but showed no reliable differences in lexicographic (length: $t(125) = 0.539, p = 0.591$; first letter: $t(125) = -0.587, p = 0.558$) or visual (color: $t(125) = -0.579, p = 0.564$; location: $t(125) = -0.346, p = 0.730$) clustering. Similarly, participants in the lexicographically ordered conditions exhibited stronger (relative to feature rich participants) lexicographic clustering (length: $t(125) = 3.426, p = 0.001$; first letter: $t(125) = 3.236, p = 0.002$) on early lists, but showed no reliable differences in semantic (category: $t(125) = -1.078, p = 0.283$; size: $t(125) = -0.310, p = 0.757$) or visual (color: $t(125) = -0.209, p = 0.835$; location: $t(125) = -0.004, p = 0.997$) clustering. And participants in the visually ordered conditions exhibited stronger visual clustering (again, relative to feature rich participants, and on early lists; color: $t(126) = 2.099, p = 0.038$; location: $t(126) = 4.392, p = 0.000$), but showed now reliable differences in semantic (category: $t(126) = 0.204, p = 0.839$; size: $t(126) = -0.093, p = 0.926$) or lexicographic (length: $t(126) = 0.714, p = 0.476$; first letter: $t(126) = 0.820, p = 0.414$) 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) while leaving clustering along other feature dimensions roughly unchanged (i.e., semantic manipulations did not affect lexicographic or color clustering, and so on).

When we closely examined the sequences of words participants recalled in 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 striking 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

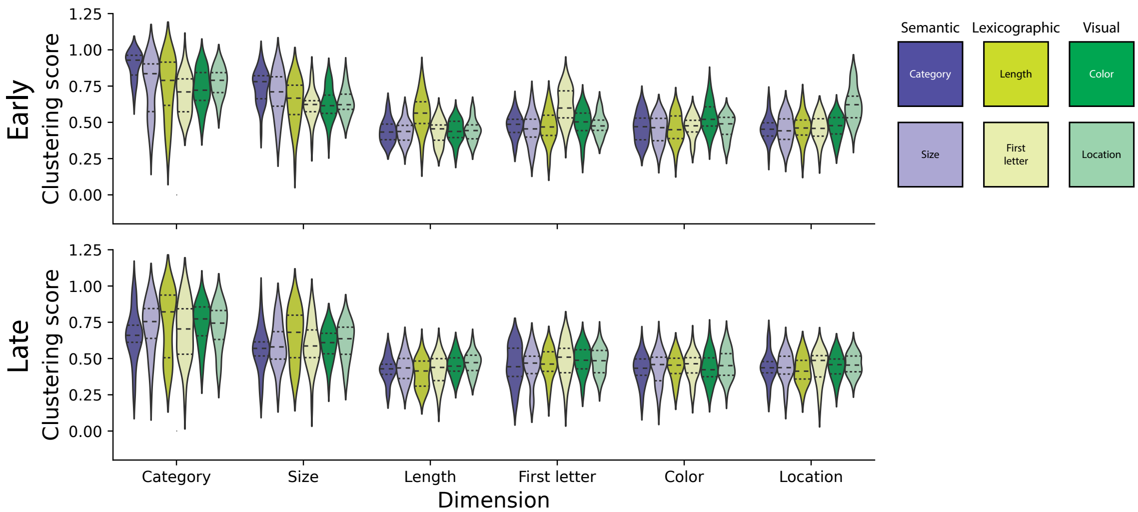

Figure 4: Memory "fingerprints" (order manipulation conditions). The across-participant distributions of clustering scores for each feature type (x-coordinate) are displayed for each experimental condition (color), separately for order manipulation (early, top) and randomly ordered (late, bottom) lists.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

recalls into groups of items that shared category labels. Indeed, when we compared the positions of feature changes in the study sequence (Fig. 3B; 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, the category labels changed every four items on average (dark purple peaks in Fig. 3B), 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 Fig. 3B), and participants also seemed to display an icnreased tendancy 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 idiosyncracies 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 irrelevant 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 ($t(118) = 1.936, p = 0.055$). Participants also remembered visually sorted lists better than lexicographically sorted lists ($t(119) = 2.145, p = 0.034$). However, participants showed no reliable differences in recall performance on semantically versus visually sorted lists ($t(119) = 0.113, p = 0.910$). Participants temporally clustered semantically sorted lists more strongly than either lexicographically ($t(118) = 5.572, p < 0.001$) or visually ($t(119) = 6.215, p < 0.001$) sorted lists, but did not show reliable differences in temporal clustering on lexicographically versus visually sorted lists ($t(119) = 0.189, p = 0.850$). Participants also showed reliably more semantic clustering on semantically sorted lists than lexicographically (category: $t(118) = 3.492, p = 0.001$, size: $t(118) = 3.972, p < 0.001$) or visually (category: $t(119) = 2.702, p = 0.008$, size: $t(119) = 4.230, p < 0.001$) sorted lists; more lexicographic clustering on lexicographically sorted lists than semantically (length: $t(118) = 3.112, p = 0.002$; first letter: $t(118) = 3.686, p = 0.000$) or visually (length: $t(119) = 3.024, p = 0.003$; first letter: $t(119) = 2.644, p = 0.009$) sorted lists; and more visual clustering on visually sorted lists than semantically (color: $t(119) = -2.659, p = 0.009$; location: $t(119) = -4.604, p = 0.000$) or lexicographically (color: $t(119) = -2.366, p = 0.020$; location: $t(119) = -4.265, p < 0.001$) sorted lists. In summary, sorting lists by different features appeared to have slightly different effects on overall memory performance and temporal clustering, and people 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., effect 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: $t(125) = 0.487, p = 0.627$; lexicographic: $t(125) = 0.878, p = 0.382$; visual: $t(126) = 1.437, p = 0.153$). Nor did we observe any reliable differences in temporal clustering on late lists (relative to late feature rich lists; semantic: $t(125) = 0.146, p = 0.884$; lexicographic: $t(125) = 0.923, p = 0.358$; visual: $t(126) = 0.525, p = 0.601$). 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; $t(126) = 2.199, p = 0.030$), we observed no reliable differences in any type of feature clustering on late order manipulation condition lists versus late feature rich lists

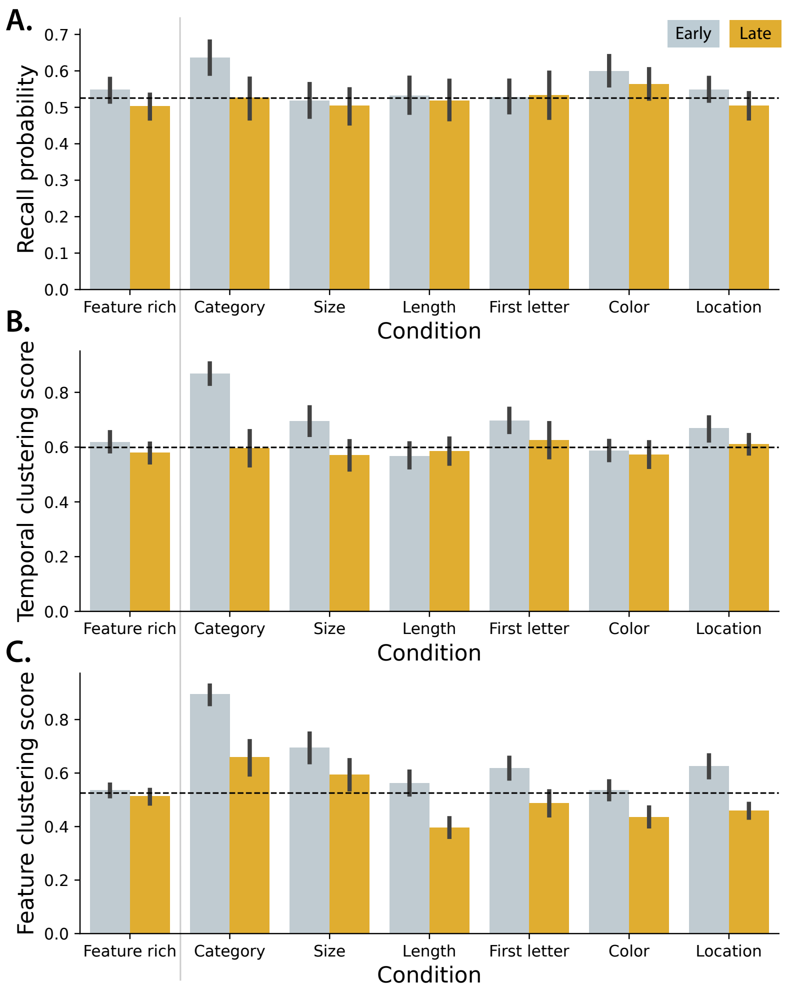

Figure 5: Recall probability and clustering scores on early and late lists. The bar heights display the average (across participants) recall probabilities (**A.**), temporal clustering scores (**B.**), and feature clustering scores (**C.**) for early (gray) and late (gold) lists. For the feature rich bars (left), the feature clustering scores are averaged across features. 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.

($||t||$s $\leq 1.234, p$s ≥ 0.220).

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. We defined participants' *memory fingerprints* as the set of temporal and feature clustering scores. A participant's memory fingerprint describes how they tend to retrieve memories of the studied items, perhaps searching 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 on 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 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 visiual 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).

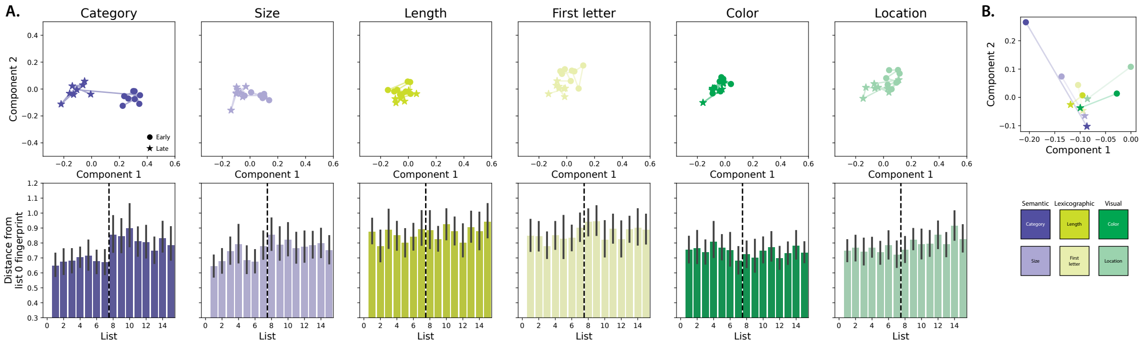

Figure 6: Memory fingerprint dynamics (order manipulation conditions). A. Each column (and color) reflects an experimental condition. In the top panels, each marker displays a 2D projection of the (across-participant) average memory fingerprint for one list. Order manipulation (early) lists are denoted 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 list 0 fingerprint, 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 (control) conditions.

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: $r(179) = 0.537, p < 0.001$; late: $r(179) = 0.492, p = 0.000$) as well as for each condition individually for early (rs ≥ 0.386, all ps ≤ 0.035) and late (rs ≥ 0.462, all ps ≤ 0.010) 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: $r(4) = 0.526, p = 0.284$; late: $r(4) = -0.257, p = 0.623$; see insets of panels A and B). We observed carryover associations between feature clustering and recall performance

(Fig. 7C, D). Participants who showed stronger feature clustering on early lists tended to recall more items on late lists (across conditions: $r(179) = 0.492, p < 0.001$; all conditions individually: rs ≥ 0.462, all ps ≤ 0.010). Participants who recalled more items on early lists also tended to show stronger feature clustering on late lists (across conditions: $r(179) = 0.280, p < 0.001$; all non-visual conditions: rs ≥ 0.445, all ps ≤ 0.014; color: $r(29) = 0.298, p = 0.103$; location: $r(28) = 0.354, p = 0.055$). Neither of these effects showed condition-level trends (early feature clustering versus late recall probability: $r(4) = -0.299, p = 0.565$; early recall probability versus late feature clustering: $r(4) = 0.400, p = 0.432$). 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. **??**E), this trend held overall ($r(179) = 0.924, p < 0.001$), for each condition individually (all rs ≥ 0.822, all ps < 0.001), and across conditions ($r(4) = 0.964, p = 0.002$). For late lists (Fig. **??**F), the results were more variable (overall: $r(179) = 0.348, p = 0.000$; all non-visual conditions: rs ≥ 0.382, all ps ≤ 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 trended towards showing stronger temporal clustering on later lists (overall: $r(179) = 0.301, p < 0.001$; for individual conditions: all rs ≥ 0.297, all ps ≤ 0.111; across conditions: $r(4) = 0.107, p = 0.840$). And participants who showed stronger temporal clustering on early lists trended towards showing stronger feature clustering on later lists (overall: $r(179) = 0.579, p < 0.001$; all non-visual conditions: rs ≥ 0.323, all ps ≤ 0.082; visual conditions: rs ≥ 0.089, all ps ≤ 0.632; across conditions: $r(4) = 0.916, p = 0.010$). Taken together, the results displayed in Figure 7 show that

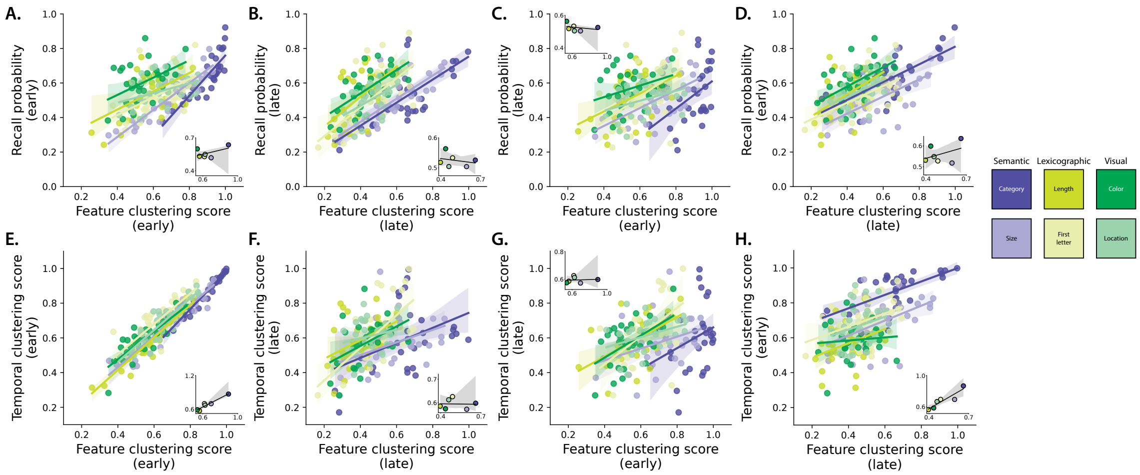

Figure 7: Interactions between feature clustering, recall probability, and contiguity. A. Recall probability versus feature clustering scores for order manipulation (early) lists. **B.** Recall probability versus feature clustering for randomly ordered (late) lists. **C.** Recall probability on late 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.

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: $r(179) = 0.592, p < 0.001$; non-visual feature conditions: all rs $\geq$ 0.350, all ps $\leq$ 0.058; color: $r(29) = -0.071, p = 0.704$; location: $r(28) = 0.032, p = 0.868$; across conditions: $r(4) = 0.934, p = 0.006$). 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 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 larger improvement in recall (Fig. 8B; overall: $r(179) = 0.378, p < 0.001$; category: $r(28) = 0.419, p = 0.021$; size: $r(28) = 0.737, p < 0.001$; non-semantic conditions: all rs $\leq$ 0.252, all ps $\geq$ 0.179; across conditions: $r(4) = 0.773, p = 0.072$) on late lists, relative to early lists. 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: $r(179) = 0.434, p < 0.001$; category: $r(28) = 0.229, p = 0.223$; all non-category conditions: all rs $\geq$ 0.448, all ps $\leq$ 0.012; across conditions: $r(4) = 0.598, p = 0.210$). We suggest two potential interpretations of these findings. First, it is possible that some participants are more "maliable" 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

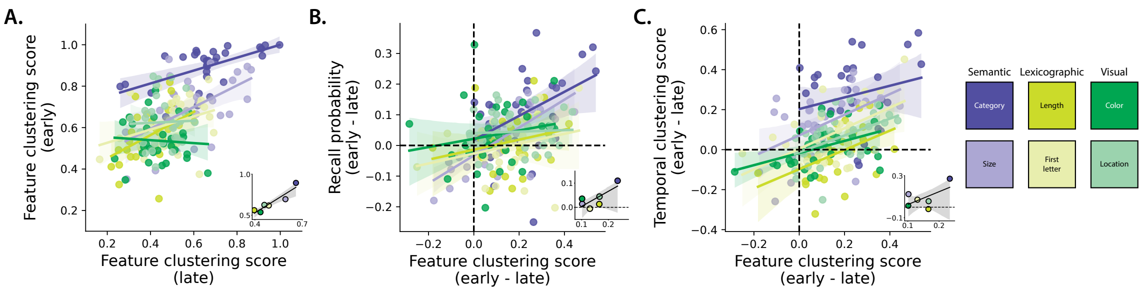

Figure 8: Feature clustering carryover effects. **A.** Feature clustering scores for ordder manipulation (early) versus randomly ordered (late) lists. **B.** Accuracy differences (on early versus late lists) versus feature clustering "carryover" (defined as the differences between the average clustering 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.

participants might demonstrate strong recall performance not because of their inherantly superior memory abilities, but rather because the specific condition they were assigned to happened to be especially easy for them, given their pre-experimental tendancies.

Figure S3.

Figure S7.

Figure S4.

Discussion

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

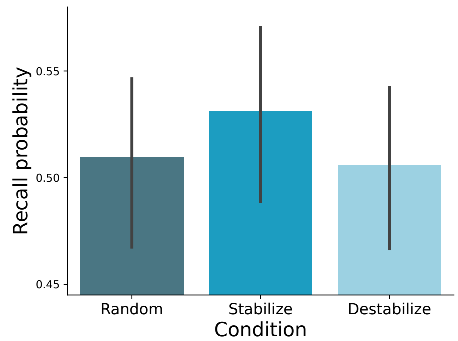

Figure 9: Recall performance (adaptive conditons). The bars display the average probability of recall (taken across words, lists, and participants) for lists from each adaptive condition. Error bars denote bootstrap-estimated 95% confidence intervals. For additional details about participants' behavior and performance during the adaptive conditions, see Figure S2.

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.

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.

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.

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

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.

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.

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.

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., 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.

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

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

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.

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

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., 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., 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.

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.

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.

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.

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.

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.

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.

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

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

Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclopedic Refernce*, 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.

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

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.

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

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.

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

Xu, X., Zhu, Z., and Manning, J. R. (2022). 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.

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.