

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

3 Jeremy R. Manning^{1,*}, Emily C. Whitaker¹, Paxton C. Fitzpatrick¹,
Madeline R. Lee¹, Allison M. Frantz¹, Bryan J. Bollinger¹,
Darya Romanova¹, Campbell E. Field¹, and Andrew C. Heusser^{1,2}

¹Dartmouth College

²Akili Interactive Labs

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

4 **Abstract**

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

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

19 Introduction

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

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

35 Our understanding of how we form situation models and schemas, and how they in-
36 teract with our subjective experiences and memories, is constrained in part by substantial
37 differences in how we study these processes. Situation models and schemas are most often
38 studied using “naturalistic” stimuli such as narratives and movies (Nastase et al., 2020;
39 Zwaan et al., 1995; Zwaan and Radvansky, 1998). In contrast, our understanding of how
40 we organize our memories has been most widely informed by more traditional paradigms
41 like free recall of random word lists (Kahana, 2012, 2020). In free recall paradigms, partic-
42 ipants study lists of items and are instructed to recall the items in any order they choose.
43 The orders in which words come to mind can provide insights into how participants have

44 organized their memories of the studied words. Because random word lists are unstruc-
45 tured by design, it is not clear if or how non-trivial situation models might apply to these
46 stimuli. As we unpack below, this provides an important motivation for our current study,
47 which uses free recall of *structured* lists to help bridge the gap between these two lines of
48 research.

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

61 Over the past half-century, context-based models have had impressive success at ex-
62 plaining many stereotyped behaviors observed during free recall and other list-learning
63 tasks (Estes, 1955; Glenberg et al., 1983; Howard and Kahana, 2002a; Kimball et al., 2007;
64 Polyn and Kahana, 2008; Polyn et al., 2009; Raaijmakers and Shiffrin, 1980; Sederberg
65 et al., 2008; Shankar and Howard, 2012; Sirotin et al., 2005). These phenomena include
66 the well-known recency and primacy effects (superior recall of items from the end and, to
67 a lesser extent, from the beginning of the studied list), as well as semantic and temporal
68 clustering effects (Howard and Kahana, 2002b; Kahana et al., 2008). The contiguity effect

69 is an example of temporal clustering, which is perhaps the dominant form of organization
70 in free recall. This effect can be seen in people’s tendencies to successively recall items that
71 occupied neighboring positions in the studied list (Kahana, 1996). There are also striking
72 effects of semantic clustering (Bousfield, 1953; Bousfield et al., 1954; Jenkins and Russell,
73 1952; Manning and Kahana, 2012; Romney et al., 1993), whereby the recall of a given
74 item is more likely to be followed by recall of a similar or related item than a dissimilar
75 or unrelated one. In general, people organize memories for words along a wide variety
76 of stimulus dimensions. According to models like the *Context Maintenance and Retrieval*
77 model (Polyn et al., 2009), the stimulus features associated with each word (e.g., the word’s
78 meaning, size of the object the word represents, letters that make up the word, font size,
79 font color, location on the screen, etc.) are incorporated into the participant’s mental con-
80 text representation (Manning, 2020; Manning et al., 2015, 2011, 2012; Smith and Vela, 2001).
81 During a memory test, any of these features may serve as a memory cue, which in turn
82 leads the participant to successively recall words that share stimulus features.

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

94 the words in a random order (Kahana, 2012).

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

117 From a theoretical perspective, we are interested in several core questions organized
118 around the central theme of how structure in our experiences affects how we remember

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

129 **Materials and methods**

130 **Participants**

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

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

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

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

155 Participants were assigned to experimental conditions based loosely on their date of
156 participation. (This aspect of our procedure helped us to more easily synchronize the ex-
157 periment databases across multiple testing computers.) Of the 490 participants who opted
158 to fill out the demographics survey, reported ages ranged from 17 to 31 years (mean: 19.1
159 years; standard deviation: 1.356 years). A total of 318 participants reported their gender
160 as female, 170 reported their gender as male, and two participants declined to report their
161 gender. A total of 442 participants reported their ethnicity as “not Hispanic or Latino,” 39
162 reported their ethnicity as “Hispanic or Latino,” and nine declined to report their ethnic-
163 ity. Participants reported their races as White (345 participants), Asian (120 participants),
164 Black or African American (31 participants), American Indian or Alaska Native (11 partic-
165 ipants), Native Hawaiian or Other Pacific Islander (four participants), Mixed race (three
166 participants), Middle Eastern (one participant), and Arab (one participant). A total of

167 five participants declined to report their race. We note that several participants reported
168 more than one of the above racial categories. Participants reported their highest degrees
169 achieved as “Some college” (359 participants), “High school graduate” (117 participants),
170 “College graduate” (seven participants), “Some high school” (five participants), “Doctor-
171 ate” (one participant), and “Master’s degree” (one participant). A total of 482 participants
172 reported no reading impairments; eight reported having mild reading impairments. A
173 total of 489 participants reported having normal color vision and one participant reported
174 having impaired color vision. A total of 482 participants reported taking no prescrip-
175 tion medications and having no recent injuries; four participants reported having ADHD,
176 one reported having dyslexia, one reported having allergies, one reported a recently torn
177 ACL/MCL, and one reported a concussion from several months prior. The participants
178 reported having consumed 0–3 cups of coffee on the day of the testing session (mean: 0.32
179 cups; standard deviation: 0.58 cups). Participants reported their current level of alertness,
180 and we converted their responses to numerical scores as follows: “very sluggish” (-2),
181 “a little sluggish” (-1), “neutral” (0), “a little alert” (1), and “very alert” (2). Across all
182 participants, the full range of alertness levels were reported (range: -2–2; mean: 0.35;
183 standard deviation: 0.89).

184 We dropped from our dataset the one participant who reported having abnormal color
185 vision, as well as 38 participants whose data were corrupted due to technical failures while
186 running the experiment or during the daily database merges. In total, this left usable data
187 from 452 participants, broken down by experimental condition as follows: feature-rich (67
188 participants), reduced (61 participants), reduced (early) (42 participants), reduced (late)
189 (41 participants), category (30 participants), size (30 participants), length (30 participants),
190 first letter (30 participants), color (31 participants), location (30 participants), and adaptive
191 (60 participants). The participant who declined to fill out their demographic survey

192 participated in the location condition, and we verified verbally that they had normal color
193 vision and no significant reading impairments.

194 **Experimental design**

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

209 **Stimuli**

210 The stimuli in our paradigm were 256 English words selected in a previous study (Ziman
211 et al., 2018). All words referred to concrete nouns and were chosen from 15 unique semantic
212 categories: body parts, building-related, cities, clothing, countries, flowers, fruits, insects,
213 instruments, kitchen-related, mammals, (US) states, tools, trees, and vegetables. We also
214 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.

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

226 We assigned the categorized words into a total of 16 lists with several constraints. First,
227 we required that each list contain exactly four unique words from each of four unique
228 categories. Second, we required that each list contain at least one word representing each
229 of the two object sizes (“small” and “large”). On average, each category was represented in
230 4.27 lists (standard deviation: 1.16 lists). Aside from these two constraints, we randomly
231 assigned each word to a single list (i.e., such that no words appeared in multiple lists
232 or were omitted entirely). After random assignment, each list contained words with an
233 average of 11.13 unique starting letters (standard deviation: 1.15 letters) and an average
234 length of 6.17 letters (standard deviation: 0.34 letters).

235 The above assignments of words to lists was performed once across all participants,
236 such that every participant studied the same set of 16 lists. In every condition, we
237 randomized the study order of these lists across participants. For participants in most
238 conditions, on some or all of the lists, we also randomly varied two additional visual
239 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 “re-

duced" 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

312 color and location distance measures always take on non-negative real values (upper-
 313 bounded at 439.94 for color, or 124.52 for location, reflecting the distances between the
 314 corresponding maximally different vectors).

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

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

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

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

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

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

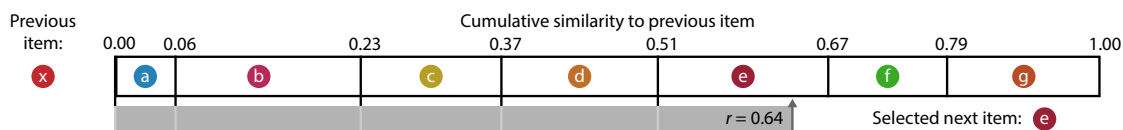


Figure 2: Generating stochastic feature-sorted lists. For a given feature dimension (e.g., color), we compute the similarity (Eqn. 1) between the feature value(s) of the previous item, x , and all yet-to-be-presented items (a – g). Next, we normalize these similarity scores so that they sum to 1. We 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.

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

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

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

365 **Feature clustering scores (uncorrected).** Feature clustering scores describe participants’
366 tendencies to recall similar presented items together in their recall sequences, where
367 “similarity” considers one given feature dimension (e.g., category, color, etc.). We based
368 our main approach to computing clustering scores on analogous temporal and semantic
369 clustering scores developed by Polyn et al. (2009). Computing the clustering score for
370 one feature dimension starts by considering the corresponding feature values from the
371 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.

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

407 Following our prior work (Heusser et al., 2017), we used a permutation-based cor-
408 rection procedure to help isolate the behavioral aspects of clustering that we were most
409 interested in. After computing the uncorrected clustering score (for the given list and
410 observed recall sequence), we constructed a “null” distribution of n additional clustering
411 scores by repeatedly randomly shuffling the order of the recalled words and recomputing
412 the clustering score for these shuffled recall sequences (we use $n = 500$ in the present
413 study). This null distribution represents an approximation of the range of clustering
414 scores one might expect to observe by “chance,” given that a hypothetical participant was
415 *not* truly clustering their recalls, but where the hypothetical participant still studied and
416 recalled exactly the same items (with the same features) as the true participant. We define
417 the *permutation-corrected clustering score* as the percentile rank of the observed uncorrected
418 clustering score in this estimated null distribution. In this way, a corrected score of 1
419 indicates that the observed score was greater than any clustering score one might expect
420 by chance—in other words, good evidence that the participant was truly clustering their

421 recalls along the given feature dimension. We applied this correction procedure to all
422 of the clustering scores (feature and temporal) reported in this paper. In Figure S4, we
423 report how participants' clustering scores along different feature dimensions (in the order
424 manipulation conditions) are correlated, and how clustering scores change across lists.

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

434 **Online "fingerprint" analysis.** The presentation orders of some lists in the adaptive con-
435 dition of our experiment (see *Adaptive condition*) were sorted according to each individual
436 participant's *current* memory fingerprint, estimated using all of the lists they had studied
437 up to that point in the experiment. Because our experiment incorporated a speech-to-text
438 component, all of the behavioral data for each participant could be analyzed just a few
439 seconds after the conclusion of the recall intervals for each list. We used the Quail Python
440 package (Heusser et al., 2017) to apply speech-to-text algorithms to the just-collected audio
441 data, aggregate the data for the given participant, and estimate the participant's memory
442 fingerprint using all of their available data up to that point in the experiment. Two aspects
443 of our implementation are worth noting: First, because memory fingerprints are com-
444 puted 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 per-

470 mutation is always equal to 1.) This yielded a “simulated fingerprint” vector \hat{f}_p for each
471 permutation p . We used these simulated fingerprints to select a specific permutation i that
472 either maximized (for stabilize lists) or minimized (for destabilize lists) the correlation
473 between \hat{f}_i and f .

474 **Computing low-dimensional embeddings of memory fingerprints**

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

492 **Factoring out the effects of temporal clustering**

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

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

515 While these temporally corrected clustering scores are useful for identifying when
516 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 $n = 1$.

541 We note that several other studies have used a slightly different approach to compute
542 these curves, by correcting for the “availability” of a given word to be recalled. For
543 example, if a participant recalls item 1, then item 2 on a given list, our approach places a
544 0 into the item 1 column for that list when computing the “probability of second recall”
545 curve. However, accounting for the fact that the participant had already recalled item
546 1, an alternative approach (e.g., Farrell, 2010) would be to count the item 1 column as
547 “unobserved” (i.e., missing data). Ultimately we chose to use the simpler variant of this
548 approach in our work, but we direct the reader to further discussion of this issue in other
549 work (Farrell, 2014; Moran and Goshen-Gottstein, 2014).

550 **Lag-conditional response probability curve**

551 The lag-conditional response probability (lag-CRP) curve (Kahana, 1996) reflects the prob-
552 ability of recalling a given item after the just-recalled item, as a function of the items’
553 relative encoding positions (lag). In other words, a lag of 1 indicates that a recalled item
554 was presented immediately after the previously recalled item, and a lag of -3 indicates
555 that a recalled item came three items before the previously recalled item. For each re-
556 call transition (following the first recall), we computed the lag between the presentation
557 positions of the just-recalled word and the next-recalled word. We then computed the
558 proportions of transitions (between successively recalled words) for each lag, normalizing
559 for the total numbers of possible transitions. In carrying out this analysis, we excluded
560 all incorrect recalls and repetitions (i.e., recalling a word that had already appeared in
561 the current recall sequence). This yielded, for each list, a 1 by number-of-lags (-15 to
562 $+15$; 30 lags in total, excluding lags of 0) array of conditional probabilities. We averaged
563 these probabilities across lists to obtain a single lag-CRP for each participant. Because
564 transitions at large absolute lags are rare, these curves are typically displayed using range

565 restrictions (Kahana, 2012).

566 **Serial position curve**

567 Serial position curves (Murdock, 1962) reflect the proportion of participants who remember
568 each item as a function of the items' serial positions during encoding. For each participant,
569 we initialized a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then,
570 for each correct recall, we identified the presentation position of the word and entered a
571 1 into that position (row: list; column: presentation position) in the matrix. This resulted
572 in a matrix whose entries indicated whether or not the words presented at each position,
573 on each list, were recalled by the participant (depending on whether the corresponding
574 entries were set to 1 or 0). Finally, we averaged over the rows of the matrix to yield a
575 1 by 16 array representing the proportion of words at each position that the participant
576 remembered.

577 **Identifying event boundaries**

578 We used the distances between feature values for successively presented words (see *Defin-*
579 *ing feature-based distances*) to estimate "event boundaries" where the feature values changed
580 more than usual (DuBrow and Davachi, 2016; Ezzyat and Davachi, 2011; Manning et al.,
581 2016; Radvansky and Copeland, 2006; Swallow et al., 2011, 2009). For each list, for each
582 feature dimension, we computed the distribution of distances between the feature values
583 for successively presented words. We defined event boundaries (e.g., Fig. 3B) as occurring
584 between any successive pair of words whose distances along the given feature dimension
585 were greater than one standard deviation above the mean for that list. Note that, because
586 event boundaries are defined for each feature dimension, each individual list may contain
587 several sets of event boundaries, each at different moments in the presentation sequence

588 (depending on the feature dimension of interest).

589 **Data and code availability**

590 All of the data analyzed in this manuscript, along with all of the code for carrying out the
591 analyses, may be found at <https://github.com/ContextLab/FRFR-analyses>. Code for run-
592 ning the non-adaptive experimental conditions may be found at <https://github.com/ContextLab/efficient-learning-code>. Code for running the adaptive experimental condition
593 may be found at <https://github.com/ContextLab/adaptiveFR>. We have also released an as-
594 sociated Python toolbox for analyzing free recall data, which may be found at [https://cdl-](https://cdl-quail.readthedocs.io/en/latest)
595 [quail.readthedocs.io/en/latest](https://cdl-quail.readthedocs.io/en/latest).

597 **Results**

598 While holding the set of words (and the assignments of words to lists) constant, we
599 manipulated two aspects of participants' experiences of studying each list. We sought to
600 understand the effects of these manipulations on participants' memories for the studied
601 words. First, we added two additional sources of visual variation to the individual word
602 presentations: font color and onscreen location. Importantly, these visual features were
603 independent of the meaning or semantic content of the words (e.g., word category, size
604 of the referent, etc.) and of the lexicographic properties of the words (e.g., word length,
605 first letter, etc.). We wondered whether this additional word-independent information
606 might facilitate recall (e.g., by providing new or richer potential ways of organizing or
607 retrieving memories of the studied words; Davachi et al., 2003; Drewnowski and Murdock,
608 1980; Hargreaves et al., 2012; Madan, 2021; Meinhardt et al., 2020; Slamecka and Barlow,
609 1979; Socher et al., 2009) or impair recall (e.g., by distracting or confusing participants
610 with irrelevant information Lange, 2005; Marsh et al., 2012, 2015; Reinitz et al., 1992).

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

619 To obtain a clean preliminary estimate of the consequences on memory of randomly
 620 varying the font colors and locations of presented words (versus holding the font color
 621 fixed at black, and the words’ locations fixed at the center of the screen), we compared par-
 622 ticipants’ performance on the *feature-rich* and *reduced* experimental conditions (see *Random*
 623 *conditions*, Fig. S1). In the feature-rich condition, the words’ colors and locations varied
 624 randomly, and in the reduced condition, words were always presented in black, at the
 625 center of the display. Aggregating across all lists for each participant, we found no differ-
 626 ence in recall accuracy (i.e., the proportions of words successfully recalled) for feature-rich
 627 versus reduced lists ($t(126) = -0.290$, $p = 0.772$, Cohen’s $d = -0.051$, bootstrap-estimated
 628 95% confidence interval (CI) = $[-2.387, 1.768]$). However, participants in the feature-rich
 629 condition clustered their recalls substantially more along every dimension we examined
 630 (temporal clustering: $t(126) = 10.632$, $p < 0.001$, $d = 1.882$, CI = $[7.786, 14.386]$; seman-
 631 tic category clustering: $t(126) = 10.148$, $p < 0.001$, $d = 1.796$, CI = $[7.324, 13.778]$; size
 632 clustering: $t(126) = 12.033$, $p < 0.001$, $d = 2.129$, CI = $[9.030, 15.918]$; word length clus-
 633 tering: $t(126) = 10.720$, $p < 0.001$, $d = 1.897$, CI = $[7.442, 15.174]$; first letter clustering:
 634 $t(126) = 6.679$, $p < 0.001$, $d = 1.182$, CI = $[4.490, 9.611]$; see *Permutation-corrected feature*
 635 *clustering scores* for more information about how we quantified each participant’s clus-

636 tering tendencies.) Taken together, these comparisons suggest that adding new features
637 changes how participants organize their memories of studied words, even when those
638 new features are independent of the words themselves and vary randomly across words.
639 We found no evidence that those additional uninformative features were distracting (in
640 terms of their impact on memory performance), but they did affect participants' recall
641 dynamics (measured via their clustering scores).

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

661 pants' fingerprints are relatively stable across lists, and that each participant's fingerprint
662 is unique to them.

663 We next asked whether adding these incidental visual features to later lists (after
664 the participants had already studied impoverished lists), or removing the visual features
665 from later lists (after the participants had already studied visually diverse lists) might
666 affect memory performance. In other words, we sought to test for potential effects of
667 changing the "richness" of participants' experiences over time. All participants studied
668 and recalled a total of 16 lists; we defined *early* lists as the first eight lists and *late* lists
669 as the last eight lists each participant encountered. To help interpret our results, we
670 compared participants' memories on early versus late lists in the above feature-rich and
671 reduced conditions. Participants in both conditions remembered more words on early
672 versus late lists (feature-rich: $t(66) = 4.553$, $p < 0.001$, $d = 0.233$, $CI = [2.427, 7.262]$;
673 reduced: $t(60) = 2.434$, $p = 0.018$, $d = 0.134$, $CI = [0.493, 4.910]$). Participants in the
674 feature-rich (but not reduced) conditions exhibited more temporal clustering on early
675 versus late lists (feature-rich: $t(66) = 2.268$, $p = 0.027$, $d = 0.181$, $CI = [0.437, 4.425]$;
676 reduced: $t(60) = 0.986$, $p = 0.328$, $d = 0.061$, $CI = [-0.897, 3.348]$). And participants
677 in both conditions tended to exhibit more semantic clustering on early versus late lists
678 (feature-rich, category: $t(66) = 3.684$, $p < 0.001$, $d = 0.220$, $CI = [1.733, 5.732]$; feature-
679 rich, size: $t(66) = 1.629$, $p = 0.108$, $d = 0.100$, $CI = [-0.207, 3.905]$; reduced, category:
680 $t(60) = 2.755$, $p = 0.008$, $d = 0.177$, $CI = [0.761, 5.189]$; reduced, size: $t(60) = 3.081$, $p =$
681 0.003 , $d = 0.201$, $CI = [1.210, 5.326]$). Participants in the reduced (but not feature-
682 rich) conditions tended to exhibit more lexicographic clustering on early versus late lists
683 (feature-rich, word length: $t(66) = -0.100$, $p = 0.921$, $d = -0.010$, $CI = [-2.217, 1.899]$;
684 feature rich, first letter: $t(66) = 0.412$, $p = 0.681$, $d = 0.045$, $CI = [-1.645, 2.461]$; reduced,
685 word length: $t(60) = 3.762$, $p < 0.001$, $d = 0.261$, $CI = [1.604, 6.821]$; reduced, first

letter: $t(60) = 1.721$, $p = 0.090$, $d = 0.175$, $CI = [-0.138, 4.098]$). 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): $t(41) = 1.499$, $p = 0.141$, $d = 0.098$, $CI = [-0.345, 3.579]$; reduced (late): $t(40) = 1.462$, $p = 0.152$, $d = 0.121$, $CI = [-0.376, 2.993]$), temporal clustering (reduced (early): $t(41) = 0.857$, $p = 0.396$, $d = 0.068$, $CI = [-1.012, 2.896]$; reduced (late): $t(40) = 1.244$, $p = 0.221$, $d = 0.128$, $CI = [-0.894, 3.088]$), nor feature-based clustering (reduced (early), category: $t(41) = 0.707$, $p = 0.484$, $d = 0.068$, $CI = [-1.314, 2.830]$; reduced (early), size: $t(41) = 0.803$, $p = 0.427$, $d = 0.079$, $CI = [-1.142, 2.953]$; reduced (early), length: $t(41) = 0.461$, $p = 0.648$, $d = 0.060$, $CI = [-1.545, 2.462]$; reduced (early), first letter: $t(41) = 0.781$, $p = 0.439$, $d = 0.101$, $CI = [-1.039, 2.881]$; reduced (late), category: $t(40) = 0.101$, $p = 0.920$, $d = 0.009$, $CI = [-1.776, 2.307]$; reduced (late), size: $t(40) = 0.555$, $p = 0.582$, $d = 0.058$, $CI = [-1.444, 2.274]$; reduced (late), length:

711 $t(40) = 1.482, p = 0.146, d = 0.126, CI = [-0.444, 3.743]$; reduced (late), first letter:
712 $t(40) = -0.143, p = 0.887, d = -0.017, CI = [-2.204, 1.830]$). We hypothesized that adding
713 or removing the variability in the visual features was acting as a sort of “event boundary”
714 between early and late lists (e.g., Clewett et al., 2019; Radvansky and Copeland, 2006;
715 Radvansky and Zacks, 2017). In prior work, we (and others) have found that memories
716 formed just after event boundaries can be enhanced (e.g., due to less contextual interfer-
717 ence between pre- and post-boundary items; Flores et al., 2017; Gold et al., 2017; Manning
718 et al., 2016; Pettijohn et al., 2016).

719 We found that *adding* incidental visual features on later lists that had not been present
720 on early lists (as in the reduced (early) condition) served to enhance recall performance
721 relative to conditions where all lists had the same blends of features (accuracy for feature-
722 rich vs. reduced (early): $t(107) = -2.230, p = 0.028, d = -0.439, CI = [-4.252, -0.229]$;
723 reduced vs. reduced (early): $t(101) = -2.045, p = 0.043, d = -0.410, CI = [-3.826, 0.112]$;
724 also see Fig. S3A). However, *subtracting* irrelevant visual features on later lists that *had*
725 been present on early lists (as in the reduced (late) condition) did not appear to impact
726 recall performance (accuracy for feature-rich vs. reduced (late): $t(106) = -0.638, p =$
727 $0.525, d = -0.126, CI = [-2.720, 1.362]$; reduced vs. reduced (late): $t(100) = -0.407, p =$
728 $0.685, d = -0.082, CI = [-2.477, 1.626]$). These comparisons suggest that recall accuracy
729 has a directional component: accuracy is affected differently by removing features that had
730 initially been present versus adding features that had initially been absent. In contrast,
731 we found that participants exhibited more temporal and feature-based clustering when
732 we added incidental visual features to *any* lists (comparisons of clustering on feature-
733 rich vs. reduced lists are reported above; temporal clustering in reduced vs. reduced
734 (early) and reduced vs. reduced (late) conditions: $ts \leq -9.885, ps < 0.001$; feature-based
735 clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions:

736 $ts \leq -4.555, ps < 0.001$). Temporal and feature-based clustering were not reliably different
737 in the feature-rich, reduced (early), and reduced (late) conditions (temporal clustering
738 in feature-rich vs. reduced (early) and feature-rich vs. reduced (late) conditions: $ts \geq$
739 $-1.379, ps \geq 0.171$; feature-based clustering in feature-rich vs. reduced (early) and feature-
740 rich vs. reduced (late) conditions: $|ts| \leq 1.441, ps \geq 0.153$).

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

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

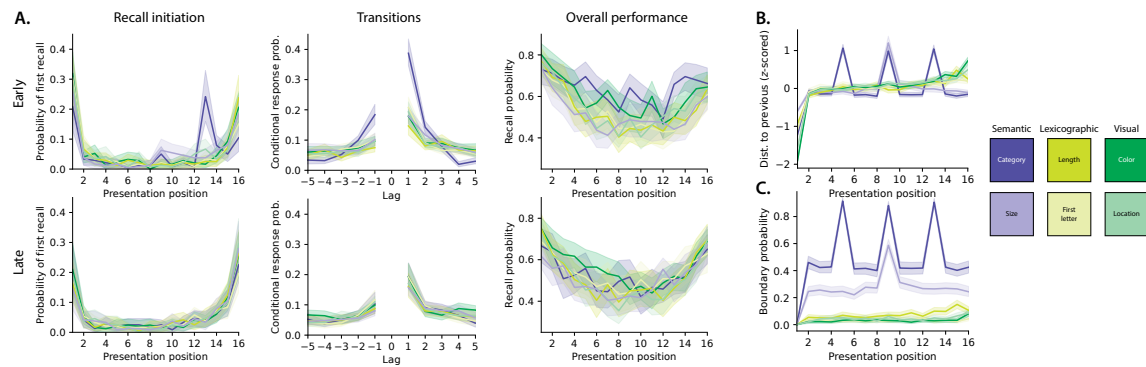


Figure 3: Recall dynamics in feature-rich free recall (order manipulation conditions). **A.** Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling each word are plotted as a function of the relative position (lag) to the word 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.** Distances between successively presented words (z-scored within condition) computed based on each condition's feature of focus, and plotted as a function of presentation position. See *Defining feature-based distances* for additional information. **C.** Proportion of event boundaries (see *Identifying event boundaries*) for each condition's feature of focus, plotted as a function of presentation position.

761 ordered condition showed an increase in memory performance on early lists (accuracy
 762 relative to early feature-rich lists; $t(95) = 3.034$, $p = 0.003$, $d = 0.667$, $CI = [1.048, 5.113]$).
 763 Participants in the color-ordered condition also showed a trending increase in memory
 764 performance on early lists (again, relative to early feature-rich lists: $t(96) = 1.850$, $p =$
 765 0.067 , $d = 0.402$, $CI = [-0.010, 3.712]$; Fig. 5A). Participants' performances on early lists in
 766 all of the other order manipulation conditions were indistinguishable from performance
 767 on the early feature-rich lists ($|t|s \leq 1.013$, $ps \geq 0.314$). Participants in both of the seman-
 768 tically ordered conditions exhibited stronger temporal clustering on early lists (vs. early
 769 feature-rich lists; category: $t(95) = 8.813$, $p < 0.001$, $d = 1.936$, $CI = [6.793, 11.751]$; size:
 770 $t(95) = 2.630$, $p = 0.010$, $d = 0.578$, $CI = [0.831, 4.866]$; Fig. 5B). Participants in the length-
 771 ordered condition tended to exhibit *less* temporal clustering on early lists relative to early
 772 feature-rich lists ($t(95) = -1.547$, $p = 0.125$, $d = -0.340$, $CI = [-3.693, 0.341]$), whereas
 773 participants in the first letter-ordered condition exhibited stronger temporal clustering on
 774 early lists ($t(95) = 2.858$, $p = 0.005$, $d = 0.628$, $CI = [1.031, 4.886]$). Participants in the
 775 visually ordered conditions exhibited more similar performance (accuracy) on early lists
 776 relative to early feature rich lists (we found a trending enhancement for participants in
 777 the color-ordered condition: $t(96) = 1.850$, $p = 0.067$, $d = 0.402$, $CI = [-0.010, 3.712]$;
 778 location: $t(95) = 0.043$, $p = 0.966$, $d = 0.010$, $CI = [-1.598, 1.729]$). Participants in
 779 the visually ordered conditions also showed similar temporal clustering on early lists
 780 relative to early feature-rich lists (color: $t(96) = -1.339$, $p = 0.184$, $d = -0.291$, $CI =$
 781 $[-3.238, 0.394]$, we found a trending increase for participants in the location-ordered
 782 condition: $t(95) = 1.705$, $p = 0.092$, $d = 0.374$, $CI = [-0.155, 3.521]$). We also com-
 783 pared feature-based clustering on early lists across the order manipulation and feature-
 784 rich conditions. Since these results were similar across both semantic conditions (cat-
 785 egory and size), both lexicographic conditions (length and first letter), and both visual

786 conditions (color and location), here we aggregate data from conditions that manipu-
 787 lated each of these three feature groupings in our comparisons, to simplify the pre-
 788 sentation. On early lists, participants in the semantically ordered conditions exhibited
 789 stronger semantic clustering relative to participants in the feature-rich condition (category:
 790 $t(125) = 2.722$, $p = 0.007$, $d = 0.484$, $CI = [0.827, 4.932]$; size: $t(125) = 3.866$, $p < 0.001$, $d =$
 791 0.687 , $CI = [2.020, 5.983]$), but showed no reliable differences in lexicographic (length:
 792 $t(125) = 0.521$, $p = 0.603$, $d = 0.093$, $CI = [-1.311, 2.333]$; first letter: $t(125) = -0.842$, $p =$
 793 0.401 , $d = -0.150$, $CI = [-2.825, 1.095]$) or visual (color: $t(125) = -0.650$, $p = 0.517$, $d =$
 794 -0.116 , $CI = [-2.680, 1.249]$; location: $t(125) = -0.251$, $p = 0.802$, $d = -0.045$, $CI =$
 795 $[-2.257, 1.524]$) clustering. Similarly, participants in the lexicographically ordered con-
 796 ditions exhibited stronger (relative to feature rich participants) lexicographic clustering
 797 (length: $t(125) = 3.682$, $p < 0.001$, $d = 0.655$, $CI = [1.890, 5.569]$; first letter: $t(125) =$
 798 5.134 , $p < 0.001$, $d = 0.912$, $CI = [3.251, 7.258]$) on early lists, but showed no reliable differ-
 799 ences in semantic (category: $t(125) = -1.040$, $p = 0.301$, $d = -0.185$, $CI = [-3.095, 1.092]$;
 800 size: $t(125) = 0.006$, $p = 0.995$, $d = 0.001$, $CI = [-1.933, 1.952]$) or visual (color:
 801 $t(125) = 0.092$, $p = 0.927$, $d = 0.016$, $CI = [-1.834, 1.867]$; location: $t(125) = 0.407$, $p =$
 802 0.685 , $d = 0.072$, $CI = [-1.655, 2.463]$) clustering. And participants in the visually ordered
 803 conditions exhibited stronger visual clustering (again, relative to feature-rich participants,
 804 and on early lists; color: $t(126) = 2.022$, $p = 0.045$, $d = 0.358$, $CI = [0.056, 3.965]$; location:
 805 $t(126) = 4.390$, $p < 0.001$, $d = 0.777$, $CI = [2.730, 6.199]$), but showed no reliable differ-
 806 ences in semantic (category: $t(126) = 0.012$, $p = 0.991$, $d = 0.002$, $CI = [-1.988, 1.871]$;
 807 size: $t(126) = -0.104$, $p = 0.917$, $d = -0.018$, $CI = [-2.166, 1.847]$) or lexicographic
 808 (length: $t(126) = 0.592$, $p = 0.555$, $d = 0.105$, $CI = [-1.361, 2.420]$; first letter: $t(126) =$
 809 0.040 , $p = 0.968$, $d = 0.007$, $CI = [-1.791, 1.863]$) clustering. Taken together, these order
 810 manipulation results suggest several broad patterns (Figs. 3A, 4). First, most of the order

811 manipulations we carried out did *not* reliably affect overall recall performance. Second,
812 most of the order manipulations increased participants' tendencies to temporally cluster
813 their recalls. Third, all of the order manipulations enhanced participants' clustering of
814 each condition's target feature (i.e., semantic manipulations enhanced semantic clustering,
815 lexicographic manipulations enhanced lexicographic clustering, and visual manipulations
816 enhanced visual clustering; Fig. 5C) while leaving clustering along other feature dimen-
817 sions roughly unchanged (i.e., semantic manipulations did not affect lexicographic or
818 visual clustering, and so on). Although it is not possible to fully separate feature-based
819 versus temporal clustering when considering sorted lists, we used a permutation-based
820 procedure to identify the degree of feature clustering over and above what could be ac-
821 counted for by temporal clustering alone (see *Factoring out the effects of temporal clustering*).
822 When we carried out this analysis (Fig. 5D), we found that participants exhibited more
823 semantic clustering on semantically sorted lists than on randomly ordered lists, but the
824 effects of the other order manipulations could not reliably be separated from temporal
825 clustering alone (reliable comparisons are reported in the figure).

826 When we closely examined the sequences of words participants recalled from early
827 order-manipulated lists (Fig. 3A, top panel), we noticed several differences from the dy-
828 namics of participants' recalls of randomly ordered lists (Figs. S1, S7). One difference is
829 that participants in the category condition (Fig. 3, dark purple curves) most often initiated
830 recall with the fourth-from-last item (*Recall initiation*, top left panel), whereas participants
831 who recalled randomly ordered lists tended to initiate recall with either the first or last
832 list items (Fig. S1, top left panel). We hypothesized that the participants might be "clump-
833 ing" their recalls into groups of items that shared category labels. Indeed, when we
834 compared the positions of feature changes in the study sequence (Fig. 3C; see *Identifying*
835 *event boundaries*) with the positions of items participants recalled first, we noticed a strik-

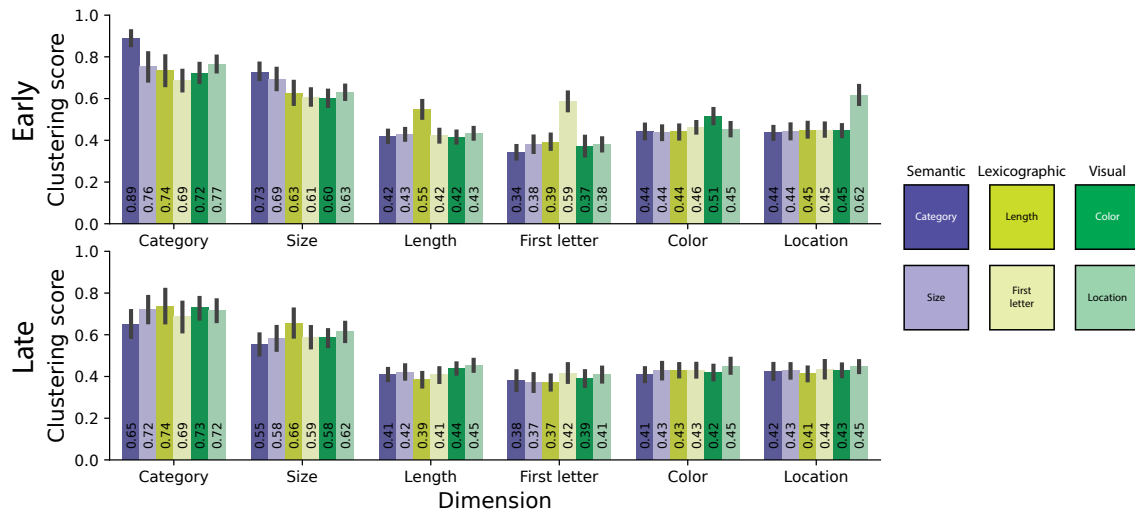


Figure 4: Memory “fingerprints” (order manipulation conditions). The across-participant average clustering scores for each feature type (*x*-axis) are displayed for each experimental condition (color), separately for order-manipulated (early, top) and randomly ordered (late, bottom) lists. Error bars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

ing correspondence in both semantic conditions. Specifically, on category-ordered lists, 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 id-

iosyncrasies 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 and 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 those sorted lexicographically ($t(118) = 1.936$, $p = 0.055$, $d = 0.353$, $CI = [0.057, 3.916]$). Participants also remembered visually sorted lists better than lexicographically sorted lists ($t(119) = 2.145$, $p = 0.034$, $d = 0.390$, $CI = [0.208, 4.254]$). However, participants showed no reliable differences in recall for semantically versus visually sorted lists ($t(119) = 0.113$, $p = 0.910$, $d = 0.021$, $CI = [-1.987, 2.097]$). Participants temporally clustered semantically sorted lists more strongly than lists sorted either lexicographically ($t(118) = 5.620$, $p < 0.001$, $d = 1.026$, $CI = [3.486, 8.010]$) or visually ($t(119) = 6.613$, $p < 0.001$, $d = 1.202$, $CI = [4.481, 9.464]$), but did not show reliable differences in temporal clustering on lexicographically versus visually sorted lists ($t(119) = 0.589$, $p = 0.557$, $d = 0.107$, $CI = [-1.336, 2.539]$). Participants also showed reliably more semantic clustering on semantically sorted lists than lexicographically (category: $t(118) = 3.667$, $p < 0.001$, $d = 0.670$, $CI = [1.822, 5.942]$; size: $t(118) = 3.972$, $p < 0.001$) or visually (category: $t(119) = 2.702$, $p = 0.008$; size: $t(118) = 4.043$, $p < 0.001$, $d = 0.738$, $CI = [2.145, 6.296]$) sorted lists; more lexicographic clustering on lexicographically sorted lists than semantically (length: $t(118) = 3.390$, $p < 0.001$, $d = 0.619$, $CI = [1.499, 5.661]$;

873 first letter: $t(118) = 5.705$, $p < 0.001$, $d = 1.042$, $CI = [3.841, 7.790]$) or visually (length:
874 $t(119) = 3.399$, $p < 0.001$, $d = 0.618$, $CI = [1.500, 5.527]$; first letter: $t(119) = 4.859$, $p <$
875 0.001 , $d = 0.883$, $CI = [2.860, 6.849]$) sorted lists; and more visual clustering on visu-
876 ally sorted lists than semantically (color: $t(119) = 2.673$, $p = 0.009$, $d = 0.486$, $CI =$
877 $[0.848, 4.567]$; location: $t(119) = 4.499$, $p < 0.001$, $d = 0.818$, $CI = [2.721, 6.399]$) or lex-
878 icographically (color: $t(119) = 1.988$, $p = 0.049$, $d = 0.361$, $CI = [0.102, 3.894]$; location:
879 $t(119) = 3.966$, $p < 0.001$, $d = 0.721$, $CI = [2.099, 5.862]$) sorted lists. In summary, sorting
880 lists by different features appeared to have slightly different effects on overall memory per-
881 formance and temporal clustering. Participants also tended to cluster their recalls along a
882 given feature dimension more when the studied lists were (versus were not) sorted along
883 that dimension.

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

891 At the group level, there seemed to be almost no lingering impact of sorting early lists
892 on memory for later lists. To simplify the presentation, we report these null results in
893 aggregate across the three feature groupings. Relative to memory performance on late
894 feature-rich lists, participants' memory performance in all six order manipulation condi-
895 tions showed no reliable differences (semantic: $t(125) = 0.487$, $p = 0.627$, $d = 0.087$, $CI =$
896 $[-1.661, 2.323]$; lexicographic: $t(125) = 0.878$, $p = 0.382$, $d = 0.156$, $CI = [-1.226, 3.044]$;
897 visual: $t(126) = 1.437$, $p = 0.153$, $d = 0.254$, $CI = [-0.447, 3.519]$). Nor did we observe

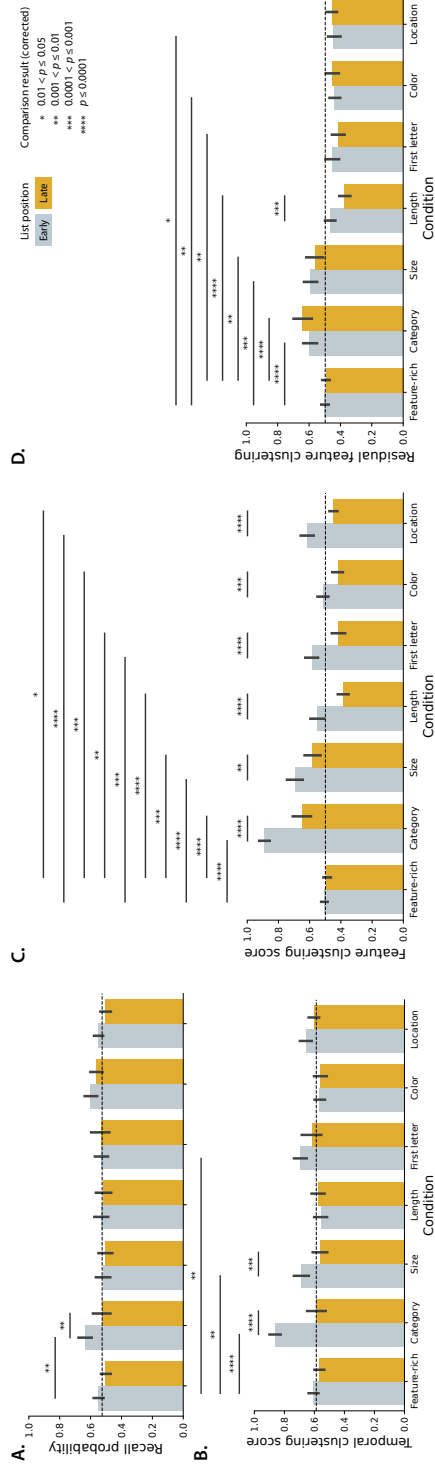


Figure 5: Recall probability and clustering scores on early and late lists. The bar heights display the average (across participants) recall probabilities (A.), temporal clustering scores (B.), feature clustering scores (C.), and residual feature clustering scores (after factoring out temporal clustering 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.

any reliable differences in temporal clustering on late lists (relative to late feature-rich lists; semantic: $t(125) = 0.157$, $p = 0.875$, $d = 0.028$, $CI = [-1.859, 1.974]$; lexicographic: $t(125) = 0.998$, $p = 0.320$, $d = 0.177$, $CI = [-0.902, 2.920]$; visual: $t(126) = 0.548$, $p = 0.585$, $d = 0.097$, $CI = [-1.450, 2.365]$). 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; $t(126) = 2.005$, $p = 0.047$, $d = 0.355$, $CI = [0.211, 3.722]$), we observed no reliable differences in any type of feature clustering on late order manipulation condition lists versus late feature-rich lists ($|t|s \leq 1.124$, $ps \geq 0.263$).

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

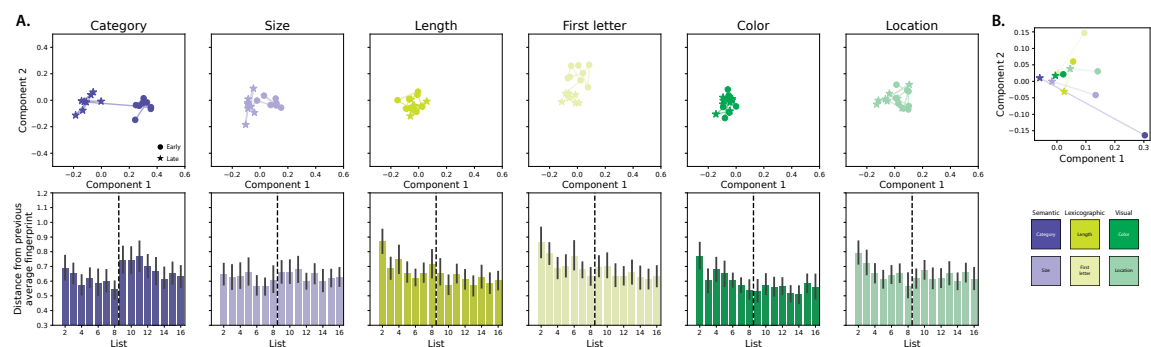


Figure 6: Memory fingerprint dynamics (order manipulation conditions). **A.** Each column (and color) reflects an experimental condition. In the top panels, each marker displays a 2D projection of the (across-participant) average memory fingerprint for a single list. Order manipulation (early) lists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

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 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: $r(179) = 0.492$, $p < 0.001$, $CI = [0.352, 0.606]$;
 late: $r(179) = 0.403$, $p < 0.001$, $CI = [0.271, 0.517]$) as well as for each condition indi-
 vidually for early ($r_s \geq 0.331$, all $p_s \leq 0.069$) and late ($r_s \geq 0.404$, all $p_s \leq 0.027$) 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 con-
 ditions where participants remembered more words (early: $r(4) = 0.511$, $p = 0.300$, $CI =$
 $[-0.999, 0.996]$; late: $r(4) = -0.304$, $p = 0.559$, $CI = [-0.833, 0.748]$; see insets of Fig. 7A
 and B). We observed carryover associations between feature clustering and recall perfor-
 mance (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:
 $r(179) = 0.230$, $p = 0.002$, $CI = [0.072, 0.372]$; all non-visual conditions individually: r_s
 ≥ 0.405 , all $p_s \leq 0.027$; color: $r(29) = 0.212$, $p = 0.251$, $CI = [-0.164, 0.532]$; location:
 $r(28) = 0.320$, $p = 0.085$, $CI = [0.011, 0.584]$). Participants who recalled more items on
 early lists also tended to show stronger feature clustering on late lists (across conditions:
 $r(179) = 0.464$, $p < 0.001$, $CI = [0.321, 0.582]$; individual conditions: all $r_s \geq 0.377$, all p_s
 ≤ 0.040). Neither of these effects showed condition-level trends (early feature clustering
 versus late recall probability: $r(4) = -0.338$, $p = 0.512$, $CI = [-0.971, 0.634]$; early recall
 probability versus late feature clustering: $r(4) = 0.451$, $p = 0.369$, $CI = [-0.986, 0.998]$). We
 also looked for associations between feature clustering and temporal clustering. Across
 every order manipulation condition, participants who exhibited stronger feature cluster-
 ing also exhibited stronger temporal clustering. For early lists (Fig. 7E), this trend held
 overall ($r(179) = 0.916$, $p < 0.001$, $CI = [0.893, 0.936]$), for each condition individually
 (all $r_s \geq 0.822$, all $p_s < 0.001$), 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: $r_s \geq 0.382$, all $p_s \leq 0.037$; color: $r(29) = 0.453$, $p = 0.011$; location:

960 $r(28) = 0.190$, $p = 0.314$; across-conditions: $r(4) = -0.036$, $p = 0.945$). While less ro-
 961 bust than the carryover associations between feature clustering and recall performance,
 962 we also observed some carryover associations between feature clustering and temporal
 963 clustering (Fig. 7G, H). Participants who showed stronger feature clustering on early lists
 964 showed stronger temporal clustering on later lists (overall: $r(179) = 0.464$, $p < 0.001$, CI =
 965 $[0.321, 0.582]$; for individual conditions: all $r_s \geq 0.377$, all $p_s \leq 0.040$; across conditions:
 966 $r(4) = 0.451$, $p = 0.369$, CI = $[-0.986, 0.998]$). And participants who showed stronger
 967 temporal clustering on early lists trended towards showing stronger feature clustering on
 968 later lists (overall: $r(179) = 0.266$, $p < 0.001$, CI = $[0.129, 0.396]$; for individual conditions:
 969 all $r_s \geq 0.298$, all $p_s \leq 0.110$; across conditions: $r(4) = 0.064$, $p = 0.903$, CI = $[-0.972, .$
 970 Taken together, the results displayed in Figure 7 show that participants who were more
 971 sensitive to the order manipulations (i.e., participants who showed stronger feature clus-
 972 tering for their condition's feature on early lists) remembered more words and showed
 973 stronger temporal clustering. These associations also appeared to carry over across lists,
 974 even when the items on later lists were presented in a random order.

975 If participants show different sensitivities to order manipulations, how do their be-
 976 haviors carry over to later lists? We found that participants who showed strong feature
 977 clustering on early lists often tended to show strong feature clustering on late lists (Fig. 8A;
 978 overall across participants and conditions: $r(179) = 0.591$, $p < 0.001$, CI = $[0.472, 0.682]$;
 979 category: $r(28) = 0.590$, $p < 0.001$, CI = $[0.354, 0.756]$; size: $r(28) = 0.488$, $p =$
 980 0.006 , CI = $[0.134, 0.732]$; length: $r(28) = 0.384$, $p = 0.036$, CI = $[0.040, 0.681]$; first
 981 letter: $r(28) = 0.202$, $p = 0.284$, CI = $[-0.273, 0.620]$; color: $r(29) = -0.183$, $p = 0.325$, CI =
 982 $[-0.562, 0.258]$; location: $r(28) = 0.031$, $p = 0.870$, CI = $[-0.240, 0.296]$; across condi-
 983 tions: $r(4) = 0.942$, $p = 0.005$, CI = $[0.442, 1.000]$). Although participants tended to
 984 show weaker feature clustering on late lists (Fig. 6) on *average*, the associations between

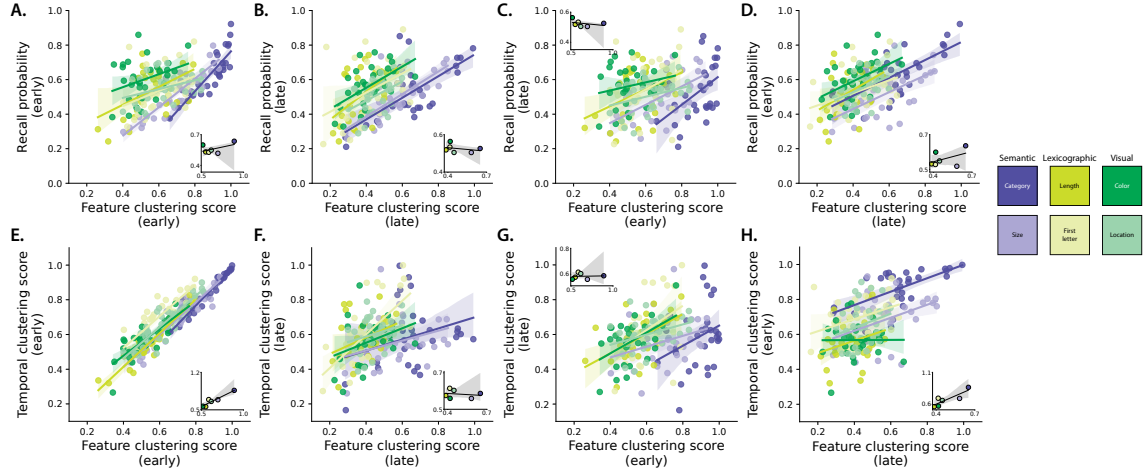


Figure 7: Interactions between feature clustering, recall probability, and contiguity. A. Recall probability versus feature clustering scores for order manipulation (early) lists. B. Recall probability versus feature clustering for randomly ordered (late) lists. C. Recall probability on late lists versus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

985 early and late lists for individual participants suggests that some influence of early or-
 986 der manipulations may linger on late lists. We found that participants who exhibited
 987 larger carryover in feature clustering (i.e., continued to show strong feature clustering
 988 on late lists) for the semantic order manipulations (but not other manipulations) also
 989 tended to show a smaller decrease in recall on early versus late lists (Fig. 8B; overall:
 990 $r(179) = 0.307$, $p < 0.001$, $CI = [0.148, 0.469]$; category: $r(28) = 0.350$, $p = 0.058$, $CI =$
 991 $[0.050, 0.642]$; size: $r(28) = 0.708$, $p < 0.001$, $CI = [0.472, 0.862]$; length: $r(28) = 0.205$, $p =$
 992 0.276 , $CI = [-0.109, 0.492]$; first letter: $r(28) = 0.081$, $p = 0.672$, $CI = [-0.433, 0.597]$;
 993 color: $r(29) = 0.155$, $p = 0.406$, $CI = [-0.174, 0.541]$; location: $r(28) = 0.052$, $p =$
 994 0.787 , $CI = [-0.307, 0.360]$; across conditions: $r(4) = 0.635$, $p = 0.176$, $CI = [-0.924, 0.981]$.
 995 Participants who exhibited larger carryover in feature clustering also tended to show
 996 stronger temporal clustering on late lists (relative to early lists) for all but the category
 997 condition (Fig. 8C; overall: $r(179) = 0.426$, $p < 0.001$, $CI = [0.285, 0.544]$; category:
 998 $r(28) = 0.110$, $p = 0.564$, $CI = [-0.284, 0.442]$; all non-category conditions: all $rs \geq 0.406$,
 999 all $ps \leq 0.023$; across conditions: $r(4) = 0.649$, $p = 0.163$, $CI = [-0.856, 0.988]$).

1000 We suggest two potential interpretations of these findings. First, it is possible that
 1001 some participants are more “malleable” or “adaptable” with respect to how they organize
 1002 incoming information. When presented with list of items sorted along *any* feature dimen-
 1003 sion, they will simply adopt that feature as a dominant dimension for organizing those
 1004 items and subsequent (randomly ordered) items. This flexibility in memory organization
 1005 might afford such participants a memory advantage, explaining their strong recall perfor-
 1006 mance. An alternative interpretation is that each participant comes into our study with a
 1007 “preferred” way of organizing incoming information. If they happen to be assigned to an
 1008 order manipulation condition that matches their preferences, then they will appear to be
 1009 “sensitive” to the order manipulation and also exhibit a high degree of carryover in feature

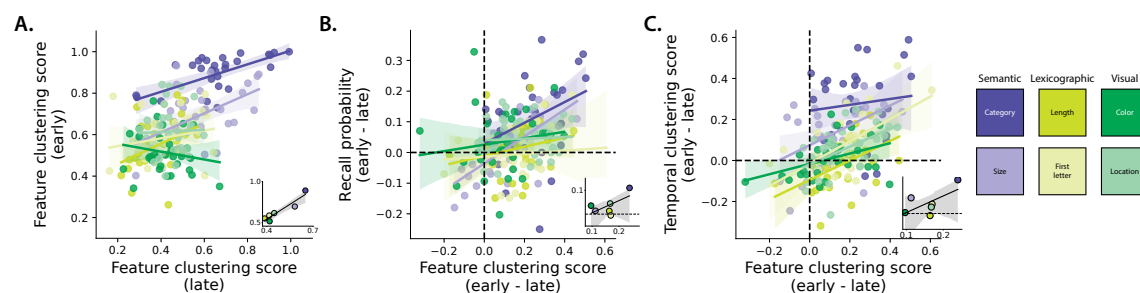


Figure 8: Feature clustering carryover effects. **A.** Feature clustering scores for order manipulation (early) versus randomly ordered (late) lists. **B.** Accuracy differences (on early versus late lists) versus feature clustering “carryover” (defined as the differences between the average clustering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

1010 clustering from early to late lists. These participants might demonstrate strong recall per-
 1011 formance not because of their inherently superior memory abilities, but rather because the
 1012 specific condition they were assigned to happened to be especially easy for them, given
 1013 their pre-experimental tendencies. To help distinguish between these interpretations, we
 1014 designed an *adaptive* experimental condition (see *Adaptive condition*). The primary ma-
 1015 nipulation in the adaptive condition is that participants each experience three key types
 1016 of lists. On *random* lists, words are ordered randomly (as in the feature-rich condition).
 1017 On *stabilize* lists, the presentation order is adjusted to be maximally similar to the current
 1018 estimate of the participant’s memory fingerprint (see *Online “fingerprint” analysis*). Third,
 1019 on *destabilize* lists, the presentation order is adjusted to be *minimally* similar to the current
 1020 estimate of the participant’s memory fingerprint (see *Ordering “stabilize” and “destabilize”*
 1021 *lists by an estimated fingerprint*). The orders in which participants experienced each type
 1022 of list were counterbalanced across participants to help reduce the influence of potential
 1023 list-order effects. Because the presentation orders on stabilize and destabilize lists are

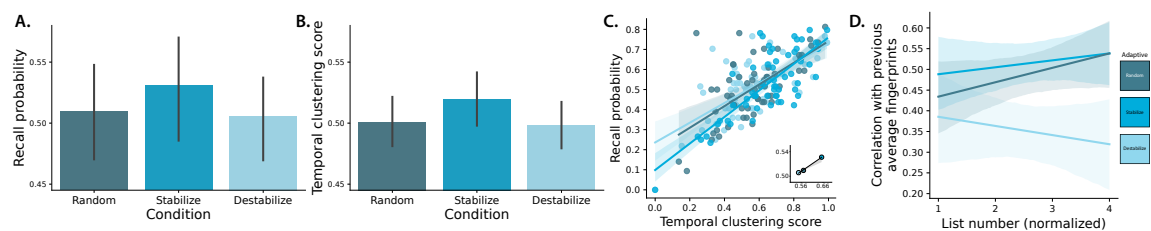


Figure 9: Adaptive free recall. **A.** Average probability of recall (taken across words, lists, and participants) for each batch of four lists in the adaptive condition. **B.** Average temporal clustering scores for each batch of lists. **C.** Recall probability versus temporal clustering scores by participant (main panel; each participant contributes one dot per batch) and averaged within batch (inset; each dot represents a single batch). **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 policy (batch) for each list. Error bars and ribbons denote bootstrap-estimated 95% confidence intervals. For additional details about participants’ behavior and performance during the adaptive condition, see Figure S2.

adjusted to best match each participant’s (potentially unique) memory fingerprint, the adaptive condition removes uncertainty about whether participants’ assigned conditions 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 ($t(59) = 1.740$, $p = 0.087$, $d = 0.095$, $CI = [-0.187, 3.761]$) and destabilize ($t(59) = 1.714$, $p = 0.092$, $d = 0.114$, $CI = [-0.351, 4.108]$) lists (Fig. 9A). Participants showed no reliable differences in their memory performance on destabilize versus random lists ($t(59) = -0.249$, $p = 0.804$, $d = -0.017$, $CI = [-2.327, 1.578]$). Participants also exhibited stronger temporal clustering on stabilize lists, relative to random ($t(59) = 3.428$, $p = 0.001$, $d = 0.306$, $CI = [1.635, 5.460]$) and destabilize ($t(59) = 4.174$, $p < 0.001$, $d = 0.374$, $CI = [1.964, 6.968]$) lists (Fig. 9B). We found no reliable differences in temporal clustering for items on random versus destabilize lists

1039 ($t(59) = -0.880$, $p = 0.382$, $d = -0.081$, $CI = [-3.165, 1.127]$).

1040 As in the other experimental manipulations, participants in the adaptive condition
1041 exhibited substantial variability with respect to their overall memory performance and
1042 their clustering tendencies (Fig. 9C). We found that individual participants who exhibited
1043 strong temporal clustering scores also tended to recall more items. This held across
1044 subjects, aggregating across all list types ($r(178) = 0.701$, $p < 0.001$, $CI = [0.590, 0.789]$),
1045 and for each list type individually (all $rs \geq 0.651$, all $ps < 0.001$). Taken together, the results
1046 from the adaptive condition suggest that each participant comes into the experiment with
1047 their own unique memory organization tendencies, as characterized by their memory
1048 fingerprint. When participants study lists whose items come pre-sorted according to their
1049 unique preferences, they tend to remember more and show stronger temporal clustering.

1050 We note that the multivariate aspect of the adaptive condition (i.e., sorting lists si-
1051 multaneously along multiple feature dimensions) provides an important contrast with
1052 the order order manipulation conditions, where we sort lists along only a single feature
1053 dimension in each condition. We found that participants “naturally” clustered their recalls
1054 along multiple feature dimensions, even when the lists they studied were not sorted along
1055 those dimensions (as in the feature-rich condition). A caveat is that the *specific* feature
1056 dimensions participants tended to cluster along varied across participants. One way to
1057 quantify the multidimensional nature of participants’ clustering tendencies is to sort each
1058 participant’s clustering scores (for each of the six feature dimensions, along with a seventh
1059 dimension to capture temporal clustering). We can then ask whether the distribution of
1060 clustering scores at each “rank” within the sorted set of scores for each participant has
1061 a mean that is reliably different from a chance value of 0.5. We carried out these tests
1062 for each set of ranked scores, and found that participants in the feature-rich condition
1063 reliably cluster their recalls along at least three dimensions, including temporal clustering

1064 (which was often ranked highest); Rank 1: $t(66) = 12.751$, $p < 0.001$, $d = 0.162$, CI =
1065 [8.702, 20.013]; Rank 2: $t(66) = 8.196$, $p < 0.001$, $d = 0.162$, CI = [4.794, 12.978]; Rank
1066 3: $t(66) = 3.243$, $p = 0.002$, $d = 0.162$, CI = [1.028, 7.051]; Rank 4: $t(66) = -3.112$, $p =$
1067 0.003 , $d = 0.162$, CI = [-5.282, -1.920]; Rank 5: $t(66) = -7.154$, $p < 0.001$, $d = 0.162$, CI =
1068 [-12.649, -5.568]; Rank 6: $t(66) = -12.608$, $p < 0.001$, $d = 0.162$, CI = [-22.114, -9.347];
1069 Rank 7: $t(66) = -18.397$, $p < 0.001$, $d = 0.162$, CI = [-27.238, -14.073].

1070 Discussion

1071 We asked participants to study and freely recall word lists. The words on each list (and
1072 the total set of lists) were held constant across participants. For each word, we considered
1073 (and manipulated) two semantic features (category and size) that reflected aspects of the
1074 *meanings* of the words, along with two lexicographic features (word length and first letter),
1075 which reflected characteristics of the words' *letters*. These semantic and lexicographic
1076 features are intrinsic to each word. We also considered and manipulated two additional
1077 visual features (color and location) that affected the *appearance* of each studied item, but
1078 could be varied independently of the words' identities. Across different experimental
1079 conditions, we manipulated how the visual features varied across words (within each
1080 list), along with the orders of each list's words. Although the participants' task (verbally
1081 recalling as many words as possible, in any order, within one minute) remained constant
1082 across all of these conditions, and although the set of words they studied from each list
1083 remained constant, our manipulations substantially affected participants' memories. The
1084 impact of some of the manipulations also affected how participants remembered *future*
1085 lists that were sorted randomly.

1086 **Recap: visual feature manipulations**

1087 We found that participants in our feature-rich condition (where we varied words' ap-
1088 pearances) recalled similar proportions of words to participants in a reduced condition
1089 (where appearance was held constant across words). However, varying the words' ap-
1090 pearances led participants to exhibit much more temporal and feature-based clustering.
1091 This suggests that even seemingly irrelevant elements of our experiences can affect how
1092 we remember them.

1093 When we held the within-list variability in participants' visual experiences fixed across
1094 lists (in the feature-rich and reduced conditions), they remembered more words from early
1095 lists than from late lists. For feature-rich lists, they also showed stronger clustering for
1096 early versus late lists. However, when we *varied* participants' visual experiences across lists
1097 (in the "reduced (early)" and "reduced (late)" conditions), these early versus late accuracy
1098 and clustering differences disappeared. Abruptly changing how incidental visual features
1099 varied across words seemed to act as a sort of "event boundary" that partially reset how
1100 participants processed and remembered post-boundary lists. Within-list clustering also
1101 increased in these manipulations, suggesting that the "within-event" words were being
1102 more tightly associated with each other.

1103 When we held the visual features constant during early lists, but then varied words'
1104 appearances in later lists (i.e., the reduced (early) condition), participants' overall memory
1105 performance improved. However, this impact was directional: when we *removed* visual
1106 features from words in late lists that had been present in early lists (i.e., the reduced (late)
1107 condition), we saw no memory improvement.

1108 **Recap: order manipulations**

1109 When we (stochastically) sorted early lists along different feature dimensions, we found
1110 several impacts on participants' memories. Sorting early lists semantically (by word cat-
1111 egory) enhanced participants' memories for those lists, but the effects on performance of
1112 sorting along other feature dimensions were inconclusive. However, each order manipu-
1113 lation substantially affected how participants *organized* their memories of words from the
1114 ordered lists. When we sorted lists semantically, participants displayed stronger semantic
1115 clustering; when we sorted lists lexicographically, they displayed stronger lexicographic
1116 clustering; and when we sorted lists visually, they displayed stronger visual clustering.
1117 Clustering along the unmanipulated feature dimensions in each of these cases was un-
1118 changed.

1119 The order manipulations we examined also appeared to induce, in some cases, a
1120 tendency to "clump" similar words within a list. This was most apparent on semantically
1121 ordered lists, where the probability of initiating recall with a given word seemed to follow
1122 groupings defined by feature change points.

1123 We also examined the impact of early list order manipulations on memory for late
1124 lists. At the group level, we found little evidence for lingering "carryover" effects of
1125 these manipulations: participants in the order manipulation conditions showed similar
1126 memory performance and clustering on late lists to participants in the corresponding
1127 control (feature-rich) condition. At the level of individual participants, however, we
1128 found several meaningful patterns.

1129 Participants who showed stronger feature clustering on early (order-manipulated) lists
1130 tended to better remember late (randomly ordered) lists. Participants who remembered
1131 early lists better also tended to show stronger feature clustering (along their condition's
1132 feature dimension) on late lists (even though the words on those late lists were presented

1133 in a random order). We also observed some (weaker) carryover effects of temporal cluster-
1134 ing. Participants who showed stronger feature clustering (along their condition's feature
1135 dimension) on early lists tended to show stronger temporal clustering on late lists. And
1136 participants who showed stronger temporal clustering on early lists also tended to show
1137 stronger feature clustering on late lists. Essentially, these order manipulations appeared to
1138 affect each participant differently. Some participants were sensitive to our manipulations,
1139 and those participants' memory performance was impacted more strongly, both for the
1140 ordered lists and for future (random) lists. Other participants appeared relatively insen-
1141 sitive to our manipulations, and those participants showed little carryover effects on late
1142 lists.

1143 These results at the individual participant level suggested to us that either (a) some
1144 participants were more sensitive to *any* order manipulation, or (b) some participants might
1145 be more (or less) sensitive to manipulations along *particular* (e.g., preferred) feature dimen-
1146 sions. To help distinguish between these possibilities, we designed an adaptive condition
1147 whereby we attempted to manipulate whether participants studied words in an order that
1148 either matched or mismatched our estimate of how they would cluster or organize the
1149 studied words in memory (i.e., their idiosyncratic memory fingerprint). We found that
1150 when we presented words in orders that were consistent with participants' memory fin-
1151 gerprints, they remembered more words overall and showed stronger temporal clustering.
1152 This comports well with the second possibility described above. Specifically, each partici-
1153 pant seems to bring into the experiment their own idiosyncratic preferences and strategies
1154 for organizing the words in their memory. When we presented the words in an order
1155 consistent with each participant's idiosyncratic fingerprint, their memory performance
1156 improved. This might indicate that the participants were spending less cognitive effort
1157 "reorganizing" the incoming words on those lists, which freed up resources to devote to

1158 encoding processes instead.

1159 **Memory consequences of feature variability**

1160 Several prior studies have examined how varying the richness or experiences, or the ex-
1161 tensive of encoding, can affect memory. Although specific details differ (Bonin et al., 2022),
1162 in general these studies have found that richer and more deeply or extensively encoded
1163 experiences are remembered better (Hargreaves et al., 2012; Madan, 2021; Meinhardt et al.,
1164 2020). Our findings help to elucidate an additional factor that may contribute to these phe-
1165 nomenon. For example, our finding that participants better remember “feature-rich” lists
1166 (where words’ appearances are varied) than “reduced” lists (where words’ appearances are
1167 held constant) only when those feature-rich lists are presented *after* reduced lists suggests
1168 that some factors that influence the richness or depth of encoding may be relative, rather
1169 than absolute. In other words, *increases* in richness (e.g., relative to a recency-weighted
1170 baseline) may be more important than the overall complexity or numbers of features.

1171 Some prior studies have suggested that people can “cue” their memories using different
1172 “strategies” or “pathways” for searching for the target information. For example, modern
1173 accounts of free recall typically posit that memory search typically begins by matching
1174 the current state of mental context with the contexts associated with other items in mem-
1175 ory (Kahana, 2020). Since context is the defining hallmark of episodic memory (Tulving,
1176 1983), context-based search can be described as an “episodic” pathway to recall. When
1177 episodic cueing fails to elicit a match, participants may then search for items that are simi-
1178 lar to the current mental context or mental state along other dimensions, such as semantic
1179 similarity (Davachi et al., 2003; Socher et al., 2009). These multiple pathways accounts of
1180 memory search also provide a potential explanation of why participants might have an
1181 easier time remembering richer stimuli (or experiences): richer stimuli and experiences

1182 might have more features that could be used to cue memory search. Our work suggests
1183 that there may be some additional factors at play with respect to the *dynamics* of these pro-
1184 cesses. In particular, we only observed memory benefits for “richer” stimuli when they
1185 were encountered after more “impoverished” stimuli (in the reduced (early) condition).
1186 This suggests that the pathways available to recall a given item may also depend on recent
1187 prior experiences.

1188 We did *not* find any evidence that changing words’ appearances *harmed* memory per-
1189 formance, e.g., by distracting them with irrelevant information (Lange, 2005; Marsh et al.,
1190 2012, 2015; Reinitz et al., 1992). Nor did we find any evidence that *changes* in the presence
1191 of potentially “distracting” features adversely affected memory. For example, when we
1192 increased or decreased the variability in words’ appearances on late versus early lists (as in
1193 the reduced (early) and reduced (late) conditions), we found no evidence that this harmed
1194 participants’ memories. One potential interpretation under the “multiple pathways to
1195 recall” framework is that the availability of multiple pathways to recall do not appear to
1196 specifically interfere with each other.

1197 **Context effects on memory performance and organization**

1198 In real-world experience, each moment’s unique blend of contextual features (where we
1199 are, who we are with, what else we are thinking of at the time, what else we experience
1200 nearby in time, etc.) plays an important role in how we interpret, experience, and re-
1201 member that moment, and how we relate it to our other experiences (e.g., for review see
1202 Manning, 2020). What are the analogues of real-world contexts in laboratory tasks like
1203 the free recall paradigm employed in our study? In general, modern formal accounts of
1204 free recall (Kahana, 2020) describe context as comprising a mix of (a) features pertaining
1205 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 (Masicampo 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-

1231 irrelevant visual features. We also found that changing the dynamics of those task-
1232 irrelevant visual features (in the reduced (early) and reduced (late) conditions) *also* affected
1233 participants' memories. This suggests that it is not only the contextual features themselves
1234 that affect memory, but also the *dynamics* of context—i.e., how the contextual features
1235 associated with each item change over time.

1236 **Priming effects on memory performance and organization**

1237 When our ongoing experiences are ambiguous, we can draw on our past experiences,
1238 expectations, and other real, perceived, or inferred cues to help resolve these ambiguities.
1239 We may also be overtly or covertly “primed” to influence how we are likely to resolve
1240 ambiguities. For example, before listening to a story with several equally plausible inter-
1241 pretations, providing participants with “background” information beforehand can lead
1242 them towards one interpretation versus another (Yeshurun et al., 2017). More broadly, our
1243 conscious and unconscious biases and preferences can influence not only how we interpret
1244 high-level ambiguities, but even how we process low-level sensory information (Katabi
1245 et al., 2023).

1246 In more simplified scenarios, like list-learning paradigms, the stimuli and tasks partic-
1247 ipants encounter before studying a given list can influence what and how they remember.
1248 For example, when participants are directed to suppress, disregard, or ignore “distracting”
1249 stimuli early on in an experiment, participants often tend to remember those stimuli less
1250 well when they are re-used as to-be-remembered targets later on in the experiment (Tip-
1251 per, 1985). In general, participants' memories can be influenced by exposing them to
1252 a wide range of positive and negative priming factors before they encounter the to-be-
1253 remembered information (Balota et al., 1992; Clayton and Chattin, 1989; Donnelly, 1988;
1254 Flexser and Tulving, 1982; Gotts et al., 2012; Huang et al., 2004; Huber, 2008; Huber et al.,

1255 2001; McNamara, 1994; Neely, 1977; Rabinowitz, 1986; Tulving and Schacter, 1991; Watkins
1256 et al., 1992; Wiggs and Martin, 1998).

1257 The order manipulation conditions in our experiment show that participants can also be
1258 primed to pick up on more subtle statistical structure in their experiences, like the dynamics
1259 of how the presentation orders of stimuli vary along particular feature dimensions. These
1260 order manipulations affected not only how participants remembered the manipulated
1261 lists, but also how they remembered *future* lists with different (randomized) temporal
1262 properties.

1263 **Free recall of blocked versus random categorized word lists**

1264 A large number of prior studies have compared participants' memories for categorized
1265 word lists that are presented in blocked versus random orders. In "blocked" lists, all
1266 of the words from a given semantic category (e.g., animals) are presented consecutively,
1267 whereas in "random" lists, the words from different categories are intermixed. Most of
1268 these studies report that participants tend to better remember blocked (versus random)
1269 lists (Bower et al., 1969; Cofer et al., 1966; D'Agostino, 1969; Dallett, 1964; Kintsch, 1970;
1270 Luek et al., 1971; Puff, 1974; Shapiro, 1970; ?; ?). Other studies suggest that these order
1271 effects may also be modulated by factors like list length and the numbers of exemplars in
1272 each category (e.g., Borges and Mangler, 1972).

1273 Although we did not directly manipulate "blocking" in our order manipulation condi-
1274 tions, our sorting procedures in those conditions (see *Constructing feature-sorted lists*) have
1275 *indirect* effects on the lists' blockiness. For example, lists that are stochastically sorted by
1276 semantic category will tend to contain runs of several same-category words in succession.
1277 Consistent with the above work on blocked versus random categorized lists, we found
1278 that participants tended to better remember lists that were sorted semantically (Fig. 5B).

1279 However, this memory improvement did not appear to extend to the other order ma-
1280 nipulation conditions we considered (e.g., to lexicographically or visually sorted lists).
1281 One possibility is that the memory benefits of blocked versus random lists are specific to
1282 semantic categories, and do not generalize to other feature dimensions. Another possi-
1283 bility is that the memory benefits are due to the presence of infrequent “jumps” between
1284 successive items (e.g., from different categories). Because the features we manipulated in
1285 the lexicographic and visual conditions were less categorical than the semantic features,
1286 feature values across words in those conditions tended to vary more gradually. Relatively
1287 stable features that are punctuated by infrequent large changes (e.g., as words transition
1288 from a same-category sequence to a new category) may also relate to perceived “event
1289 boundaries,” which can have important consequences for memory (DuBrow and Davachi,
1290 2013, 2016; DuBrow et al., 2017; Radvansky and Zacks, 2017).

1291 **Expectation, event boundaries, and situation models**

1292 Our findings that participants’ current and future memory behaviors are sensitive to
1293 manipulations in which features change over time, and how features change across items
1294 and lists, suggest parallels with studies on how we form expectations and predictions,
1295 segment our continuous experiences into discrete events, and make sense of different
1296 scenarios and situations. Each of these real-world cognitive phenomena entail identifying
1297 statistical regularities in our experiences, and exploiting those regularities to gain insight,
1298 form inferences, organize or interpret memories, and so on. Our past experiences enable
1299 us to predict what is likely to happen in the future, given what happened “next” in our
1300 previous experiences that were similar to now (Barron et al., 2020; Brigard, 2012; Chow
1301 et al., 2016; Eichenbaum and Fortin, 2009; Gluck et al., 2002; Goldstein et al., 2021; Griffiths
1302 and Steyvers, 2003; Jones and Pashler, 2007; Kim et al., 2014; Manning, 2020; Tamir and

1303 Thornton, 2018; Xu et al., 2023).

1304 When our expectations are violated, such as when our observations disagree with our
1305 predictions, we may perceive the “rules” or “situation” to have changed. *Event boundaries*
1306 denote abrupt changes in the state of our experience, for example, when we transition
1307 from one situation to another (Radvansky and Zacks, 2017; Zwaan and Radvansky, 1998).
1308 Crossing an event boundary can impair our memory for pre-boundary information and en-
1309 hance our memory for post-boundary information (DuBrow and Davachi, 2013; Manning
1310 et al., 2016; Radvansky and Copeland, 2006; Sahakyan and Kelley, 2002). Event bound-
1311 aries are also tightly associated with the notion of *situation models* and *schemas*—mental
1312 frameworks for organizing our understanding about the rules of how we and others are
1313 likely to behave, how events are likely to unfold over time, how different elements are
1314 likely to interact, and so on. For example, a situation model pertaining to a particular
1315 restaurant might set our expectations about what we are likely to experience when we
1316 visit that restaurant (e.g., what the building will look like, how it will smell when we enter,
1317 how crowded the restaurant is likely to be, the sounds we are likely to hear, etc.). Similarly,
1318 as mentioned in the *Introduction*, we might learn a schema describing how events are likely
1319 to unfold *across* any sit-down restaurant—e.g., open the door, wait to be seated, receive a
1320 menu, decide what to order, place the order, and so on. Situation models and schemas can
1321 help us to generalize across our experiences, and to generate expectations about how new
1322 experiences are likely to unfold. When those expectations are violated, we can perceive
1323 ourselves to have crossed into a new situation.

1324 In our study, we found that abruptly changing the “rules” about how the visual
1325 appearances of words are determined, or about the orders in which words are presented,
1326 can lead participants to behave similarly to what one might expect upon crossing an event
1327 boundary. Adding variability in font color and presentation location for words on late

1328 lists, after those visual features had been held constant on early lists, led participants to
1329 remember more words on those later lists. One potential explanation is that participants
1330 perceive an “event boundary” to have occurred when they encounter the first “late” list.
1331 According to contextual change accounts of memory across event boundaries (e.g., Flores
1332 et al., 2017; Gold et al., 2017; Pettijohn et al., 2016; Sahakyan and Kelley, 2002), this could
1333 help to explain why participants in the reduced (early) condition exhibited better overall
1334 memory performance. Specifically, their memory for late list items could benefit from less
1335 interference from early list items, and the contextual features associated with late list items
1336 (after the “event boundary”) might serve as more specific recall cues for those late items
1337 (relative to if the boundary had not occurred).

1338 **How do different types of clustering relate to each other, and to memory perfor-** 1339 **mance?**

1340 When the words on a studied list are presented in a random order, different types of
1341 clustering in participants’ recalls often tend to be negatively correlated. For example,
1342 words that occur nearby on the list will not (on average) tend to be semantically related, and
1343 vice versa. Therefore a participant who shows a strong tendency to temporally cluster their
1344 recalls will tend to show weaker semantic clustering, and so on (Healey and Uitvlugt, 2019;
1345 Howard and Kahana, 2002b; Sederberg et al., 2010). Further, there is some evidence that
1346 temporal clustering is positively correlated with memory performance, whereas semantic
1347 clustering is negatively correlated with memory performance (Sederberg et al., 2010).

1348 The notion of “multiple pathways to recall” discussed above (see *Memory consequences*
1349 *of feature variability*) suggests one potential explanation for these patterns. For exam-
1350 ple, temporal clustering has been proposed to reflect reliance on contextual cues in an
1351 “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

1376 more engaging than the highly simplified list-learning tasks typically employed in the
1377 psychological laboratory, perhaps leading participants to pay more attention, exert more
1378 effort, and stay more consistently motivated to perform well (Nastase et al., 2020). Another
1379 reason is that the temporal unfoldings of events and occurrences in naturalistic stimuli
1380 tend to be much more meaningful than the temporal unfoldings of items on typical lists
1381 used in laboratory memory tasks. Real-world events exhibit important associations at a
1382 broad range of timescales. For example, an early detail in a detective story may prove to
1383 be a clue to solving the mystery later on. Further, what happens in one moment typically
1384 carries some predictive information about what came before or after (Xu et al., 2023). In
1385 contrast, the lists used in laboratory memory tasks are most often ordered randomly, by
1386 design, to *remove* meaningful temporal structure in the stimulus (Kahana, 2012).

1387 On one hand, naturalistic stimuli provide a potential means of understanding how our
1388 memory systems function in the circumstances we most often encounter in our everyday
1389 lives. This implies that, to understand how memory works in the “real world,” we should
1390 study memory for stimuli that reflect the relevant statistical structure of real-world expe-
1391 riences. On the other hand, naturalistic stimuli can be difficult to precisely characterize or
1392 model, making it difficult to distinguish whether specific behavioral trends follow from
1393 fundamental workings of our memory systems, from some aspect of the stimulus, or from
1394 idiosyncratic interactions or interference between participants’ memory systems and the
1395 stimulus. This challenge implies that, to understand the fundamental nature of memory
1396 in its “pure” form, we should study memory for highly simplified stimuli that can pro-
1397 vide relatively unbiased (compared with real-world experiences) measures of the relevant
1398 patterns and tendencies.

1399 The experiment we report in this paper was designed to help bridge some of this gap
1400 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 or-

1425 ganization) these sorts of manipulations can have? Are those limits universal across
1426 people, or are there individual differences (based on prior experiences, natural strate-
1427 gies, neuroanatomy, etc.) that impose person-specific limits on the potential impact of
1428 presentation-level manipulations on memory?

1429 Our findings indicate that the ways word lists are presented affects how people re-
1430 member them. To the extent that word list memory reflects memory processes that are
1431 relevant to real-world experiences, one could imagine potential real-world applications of
1432 our findings. For example, we found that participants remembered more words when the
1433 presentation order agreed with their memory fingerprints. If analogous fingerprints could
1434 be estimated for classroom content, perhaps they could be utilized manually by teachers,
1435 or even by automated content-presentation systems, to optimize how and what students
1436 remember.

1437 **Concluding remarks**

1438 Our work raises deep questions about the fundamental nature of human learning. What
1439 are the limits of our memory systems? How much does what we remember (and how we
1440 remember) depend on how we learn or experience the to-be-remembered content? We
1441 know that our expectations, strategies, situation models learned through prior experiences,
1442 and more collectively shape how our experiences are remembered. But those aspects of
1443 our memory are not fixed: when we are exposed to the same experience in a new way, it
1444 can change how we remember that experience, and also how we remember, process, or
1445 perceive *future* experiences.

1446 **Author contributions**

1447 Conceptualization: JRM and ACH. Methodology: JRM and ACH. Software: JRM, PCF,
1448 CEF, and ACH. Analysis: JRM, PCF, and ACH. Data collection: ECW, PCF, MRL, AMF,
1449 BJB, DR, and CEF. Data curation and management: ECW, PCF, MRL, and ACH. Writing
1450 (original draft): JRM. Writing (review and editing): ECW, PCF, MRL, AMF, BJB, DR, CEF,
1451 and ACH. Supervision: JRM and ACH. Project administration: ECW and PCF. Funding
1452 acquisition: JRM.

1453 **Author note**

1454 All of the data analyzed in this manuscript, along with all of the code for carrying out the
1455 analyses may be found at <https://github.com/ContextLab/FRFR-analyses>. Code for run-
1456 ning the non-adaptive experimental conditions may be found at [https://github.com/Con-](https://github.com/ContextLab/efficient-learning-code)
1457 [textLab/efficient-learning-code](https://github.com/ContextLab/efficient-learning-code). Code for running the adaptive experimental condition
1458 may be found at <https://github.com/ContextLab/adaptiveFR>. We have also released an as-
1459 sociated Python toolbox for analyzing free recall data, which may be found at [https://cdl-](https://cdl-quail.readthedocs.io/en/latest)
1460 [quail.readthedocs.io/en/latest](https://cdl-quail.readthedocs.io/en/latest). Note that this study was not preregistered. Some of the
1461 ideas and data presented in this manuscript were also presented at the Annual Meeting
1462 of the Society for Neuroscience (2017) and the Context and Episodic Memory Symposium
1463 (2017).

1464 **Acknowledgements**

1465 We acknowledge useful discussions, assistance in setting up an earlier (unpublished)
1466 version of this study, and assistance with some of the data collection efforts from Rachel
1467 Chacko, Joseph Finkelstein, Sheherzad Mohyidin, Lucy Owen, Gal Perlman, Jake Rost,
1468 Jessica Tin, Marisol Tracy, Peter Tran, and Kirsten Ziman. Our work was supported in part

1469 by NSF CAREER Award Number 2145172 to JRM. The content is solely the responsibility
1470 of the authors and does not necessarily represent the official views of our supporting
1471 organizations. The funders had no role in study design, data collection and analysis,
1472 decision to publish, or preparation of the manuscript.

1473 **References**

- 1474 Anderson, J. R. and Bower, G. H. (1972). Recognition and retrieval processes in free recall.
1475 *Psychological Review*, 79(2):97–123.
- 1476 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its
1477 control processes. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning*
1478 *and Motivation*, volume 2, pages 89–105. Academic Press, New York, NY.
- 1479 Baddeley, A. D. (1968). Prior recall of newly learned items and the recency effect in free
1480 recall. *Canadian Journal of Psychology*, 22:157–163.
- 1481 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event
1482 schemas during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.
- 1483 Balota, D. A., Black, S. R., and Cheney, M. (1992). Automatic and attentional priming in
1484 young and older adults: reevaluation of the two-process model. *Journal of Experimental*
1485 *Psychology: Human Perception and Performance*, 18(2):485–502.
- 1486 Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a
1487 predictive coding account. *Progress in Neurobiology*, 192:101821–101834.
- 1488 Bonin, P., Thiebaut, G., Bugajska, A., and Méot, A. (2022). Mixed evidence for a richness-of-
1489 encoding account of animacy effects in memory from the generation-of-ideas paradigm.
1490 *Current Psychology*, 41:1653–1662.

- 1491 Borges, M. A. and Mangler, G. (1972). Effect of within-category spacing on free recall.
1492 *Journal of Experimental Psychology*, 92(2):207–214.
- 1493 Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged
1494 associates. *Journal of General Psychology*, 49:229–240.
- 1495 Bousfield, W. A., Sedgewick, C. H., and Cohen, B. H. (1954). Certain temporal character-
1496 istics of the recall of verbal associates. *American Journal of Psychology*, 67:111–118.
- 1497 Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive*
1498 *Psychology*, 11(2):177–220.
- 1499 Bower, G. H., Lesgold, A. M., and Tieman, D. (1969). Grouping operations in free recall.
1500 *Journal of Verbal Learning and Verbal Behavior*, 8(4):481–493.
- 1501 Brigard, F. D. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*,
1502 3(420):1–3.
- 1503 Chiu, Y.-C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., and Sahakyan, L. (2021). Sepa-
1504 ration of item and context in item-method directed forgetting. *NeuroImage*, 235:117983.
- 1505 Chow, W.-Y., Momma, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory
1506 retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.
- 1507 Clayton, K. and Chattin, D. (1989). Spatial and semantic priming effects in tests of spa-
1508 tial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
1509 15(3):495–506.
- 1510 Clewett, D., DuBrow, S., and Davachi, L. (2019). Transcending time in the brain: how
1511 event memories are constructed from experience. *Hippocampus*, 29(3):162–183.

- 1512 Cofer, C. N., Bruce, D. R., and Reicher, G. M. (1966). Clustering in free recall as a function
1513 of certain methodological variations. *Journal of Experimental Psychology: General*, 71:858–
1514 866.
- 1515 D’Agostino, P. R. (1969). The blocked-random effect in recall and recognition. *Journal of*
1516 *Verbal Learning and Verbal Behavior*, 8:815–820.
- 1517 Dallett, K. M. (1964). Number of categories and category information in free recall. *Journal*
1518 *of Experimental Psychology*, 68:1–12.
- 1519 Darley, C. F. and Murdock, B. B. (1971). Effects of prior free recall testing on final recall
1520 and recognition. *Journal of Experimental Psychology: General*, 91:66–73.
- 1521 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct
1522 medial temporal lobe processes build item and source memories. *Proceedings of the*
1523 *National Academy of Sciences, USA*, 100(4):2157–2162.
- 1524 Donnelly, R. E. (1988). Priming effects in successive episodic tests. *Journal of Experimental*
1525 *Psychology: Learning, Memory, and Cognition*, 14:256–265.
- 1526 Drewnowski, A. and Murdock, B. B. (1980). The role of auditory features in memory span
1527 for words. *Journal of Experimental Psychology: Human Learning and Memory*, 6:319–332.
- 1528 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for
1529 the sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–
1530 1286.
- 1531 DuBrow, S. and Davachi, L. (2016). Temporal binding within and across events. *Neurobi-*
1532 *ology of Learning and Memory*, 134:107–114.

- 1533 DuBrow, S., Rouhani, N., Niv, Y., and Norman, K. A. (2017). Does mental context drift or
1534 shift? *Current Opinion in Behavioral Sciences*, 17:141–146.
- 1535 Eichenbaum, H. and Fortin, N. J. (2009). The neurobiology of memory based predictions.
1536 *Philosophical Transactions of the Royal Society of London Series B*, 364(1521):1183–1191.
- 1537 Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological*
1538 *Review*, 62:145–154.
- 1539 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory?
1540 *Psychological Science*, 22(2):243–252.
- 1541 Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall:
1542 a challenge for unitary models of recency. *Journal of Experimental Psychology: Learning,*
1543 *Memory, and Cognition*, 36:324–347.
- 1544 Farrell, S. (2014). Correcting the correction of conditional recency slopes. *Psychonomic*
1545 *Bulletin and Review*, 21:1174–1179.
- 1546 Fitzpatrick, P. C., Heusser, A. C., and Manning, J. R. (2023). Text embedding models yield
1547 high-resolution insights into conceptual knowledge from short multiple-choice quizzes.
1548 *PsyArXiv*, page doi.org/10.31234/osf.io/dh3q2.
- 1549 Flexser, A. J. and Tulving, E. (1982). Priming and recognition failure. *Journal of Verbal*
1550 *Learning and Verbal Behavior*, 21:237–248.
- 1551 Flores, S., Bailey, H. R., Eisenberg, M. L., and Zacks, J. M. (2017). Event segmentation
1552 improves event memory up to one month later. *Journal of Experimental Psychology:*
1553 *Learning, Memory, and Cognition*, 43(8):1183.

- 1554 Gershman, S. J., Schapiro, A. C., Hupbach, A., and Norman, K. A. (2013). Neural context
1555 reinstatement predicts memory misattribution. *The Journal of Neuroscience*, 33(20):8590–
1556 8595.
- 1557 Glenberg, A. M., Bradley, M. M., Kraus, T. A., and Renzaglia, G. J. (1983). Studies of the
1558 long-term recency effect: support for a contextually guided retrieval theory. *Journal of*
1559 *Experimental Psychology: Learning, Memory, and Cognition*, 12:413–418.
- 1560 Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
1561 prediction” task? individual variability in strategies for probabilistic category learning.
1562 *Learning and Memory*, 9:408–418.
- 1563 Gold, D. A., Zacks, J. M., and Flores, S. (2017). Effects of cues to event segmentation on
1564 subsequent memory. *Cognitive Research: Principles and Implications*, 2(1):1.
- 1565 Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder,
1566 A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto,
1567 C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A.,
1568 Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. (2021). Thinking
1569 ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*,
1570 page doi.org/10.1101/2020.12.02.403477.
- 1571 Gotts, S. J., Chow, C. C., and Martin, A. (2012). Repetition priming and repetition sup-
1572 pression: A case for enhanced efficiency through neural synchronization. *Cognitive*
1573 *Neuroscience*, 3(3-4):227–237.
- 1574 Griffiths, T. L. and Steyvers, M. (2003). Prediction and semantic association. *Advances in*
1575 *Neural Information Processing Systems*, 15.

- 1576 Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml,
1577 M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages
1578 2338–2342.
- 1579 Hargreaves, I. S., Pexman, P. M., Johnson, J. C., and Zdrazilova, L. (2012). Richer concepts
1580 are better remembered: number of features effects in free recall. *Frontiers in Human*
1581 *Neuroscience*, 6:doi.org/10.3389/fnhum.2012.00073.
- 1582 Healey, M. K. and Uitvlugt, M. G. (2019). The role of control processes in temporal and
1583 semantic contiguity. *Memory and Cognition*, 47:719–737.
- 1584 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail:
1585 a Python toolbox for analyzing and plotting free recall data. *Journal of Open Source*
1586 *Software*, 10.21105/joss.00424.
- 1587 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal
1588 behavioral and neural signatures of transforming experiences into memories. *Nature*
1589 *Human Behavior*, 5:905–919.
- 1590 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). HyperTools: a
1591 Python toolbox for gaining geometric insights into high-dimensional data. *Journal of*
1592 *Machine Learning Research*, 18(152):1–6.
- 1593 Hogan, R. M. (1975). Interitem encoding and directed search in free recall. *Memory and*
1594 *Cognition*, 3:197–209.
- 1595 Howard, M. W. and Kahana, M. J. (1999). Contextual variability and serial position effects
1596 in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:923–
1597 941.

- 1598 Howard, M. W. and Kahana, M. J. (2002a). A distributed representation of temporal
1599 context. *Journal of Mathematical Psychology*, 46:269–299.
- 1600 Howard, M. W. and Kahana, M. J. (2002b). When does semantic similarity help episodic
1601 retrieval? *Journal of Memory and Language*, 46:85–98.
- 1602 Huang, L., Holcombe, A. O., and Pashler, H. (2004). Repetition priming in visual search:
1603 episodic retrieval, not feature priming. *Memory and Cognition*, 32:12–20.
- 1604 Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental*
1605 *Psychology: General*, 137(2):324–347.
- 1606 Huber, D. E., Shiffrin, R. M., Lyle, K. B., and Ruys, K. I. (2001). Perception and preference
1607 in short-term word priming. *Psychological Review*, 108(1):149–182.
- 1608 Isarida, T. and Isarida, T. K. (2007). Environmental context effects of background color in
1609 free recall. *Memory and Cognition*, 35(7):1620–1629.
- 1610 Jenkins, J. J. and Russell, W. A. (1952). Associative clustering during recall. *Journal of*
1611 *Abnormal and Social Psychology*, 47:818–821.
- 1612 Jones, A. C. and Pyc, M. A. (2014). The production effect: costs and benefits in free recall.
1613 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):300–305.
- 1614 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing
1615 prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 1616 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*,
1617 24:103–109.
- 1618 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York,
1619 NY.

- 1620 Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychol-*
 1621 *ogy*, 71:107–138.
- 1622 Kahana, M. J., Howard, M. W., and Polyn, S. M. (2008). Associative processes in episodic
 1623 memory. In Roediger III, H. L., editor, *Cognitive Psychology of Memory*, pages 476–490.
 1624 Elsevier, Oxford, UK.
- 1625 Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., and Yeshurun, Y. (2023). Deeper than
 1626 you think: partisanship-dependent brain responses in early sensory and motor brain
 1627 regions. *The Journal of Neuroscience*, pages doi.org/10.1523/JNEUROSCI.0895–22.2022.
- 1628 Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning
 1629 of memories by context-based prediction error. *Proceedings of the National Academy of*
 1630 *Sciences, USA*, In press.
- 1631 Kimball, D. R., Smith, T. A., and Kahana, M. J. (2007). The fSAM model of false recall.
 1632 *Psychological Review*, 114(4):954–993.
- 1633 Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 1634 Lange, E. B. (2005). Disruption of attention by irrelevant stimuli in serial recall. *Journal of*
 1635 *Memory and Language*, 43(4):513–531.
- 1636 Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases
 1637 of human memory. *Current Opinion in Behavioral Sciences*, 32:111–119.
- 1638 Lohnas, L. J., Polyn, S. M., and Kahana, M. J. (2010). Modeling intralist and interlist effects
 1639 in free recall. In *Psychonomic Society*, Saint Louis, MO.
- 1640 Luek, S. P., McLaughlin, J. P., and Cicala, G. A. (1971). Effects of blocking of input and

- 1641 blocking of retrieval cues on free recall learning. *Journal of Experimental Psychology*,
1642 91(1):159–161.
- 1643 Madan, C. R. (2021). Exploring word memorability: how well do different word properties
1644 explain item free-recall probability? *Psychonomic Bulletin and Review*, 28:583–595.
- 1645 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
1646 *Handbook of Human Memory*. Oxford University Press.
- 1647 Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
1648 function? *Psychological Review*, 128(4):711–725.
- 1649 Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. A.
1650 (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic*
1651 *Bulletin and Review*, 23(5):1534–1542.
- 1652 Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free
1653 recall. *Memory*, 20(5):511–517.
- 1654 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic
1655 memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.
- 1656 Manning, J. R., Notaro, G. M., Chen, E., and Fitzpatrick, P. C. (2022). Fitness tracking
1657 reveals task-specific associations between memory, mental health, and physical activity.
1658 *Scientific Reports*, 12(13822):doi.org/10.1038/s41598-022-17781-0.
- 1659 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory pat-
1660 terns in temporal lobe reveal context reinstatement during memory search. *Proceedings*
1661 *of the National Academy of Sciences, USA*, 108(31):12893–12897.

- 1662 Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., and Kahana, M. J. (2012).
1663 Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clus-
1664 tering during memory search. *The Journal of Neuroscience*, 32(26):8871–8878.
- 1665 Marsh, J. E., Beaman, C. P., Hughes, R. W., and Jones, D. M. (2012). Inhibitory control in
1666 memory: evidence for negative priming in free recall. *Journal of Experimental Psychology:*
1667 *Learning, Memory, and Cognition*, 38(5):1377–1388.
- 1668 Marsh, J. E., Sörqvist, P., Hodgetts, H. M., Beaman, C. P., and Jones, D. M. (2015). Distraction
1669 control processes in free recall: benefits and costs to performance. *Journal of Experimental*
1670 *Psychology: Learning, Memory, and Cognition*, 41(1):118–133.
- 1671 Masicampo, E. J. and Sahakyan, L. (2014). Imagining another context during encoding off-
1672 sets context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory,*
1673 *and Cognition*, 40(6):1772–1777.
- 1674 Masís-Obando, R., Norman, K. A., and Baldassano, C. (2022). Scheme representations in
1675 distinct brain networks support narrative memory during encoding and retrieval. *eLife*,
1676 11:e70445.
- 1677 McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental*
1678 *Psychology: Learning, Memory, and Cognition*, 20:507–520.
- 1679 Meinhardt, M. J., Bell, R., Buchner, A., and Röer, J. P. (2020). Adaptive memory: is
1680 the animacy effect on memory due to richness of encoding? *Journal of Experimental*
1681 *Psychology: Learning, Memory, and Cognition*, 46(3):416–426.
- 1682 Miller, J. F., Kahana, M. J., and Weidemann, C. T. (2012). Recall termination in free recall.
1683 *Memory and Cognition*, 40(4):540–550.

- 1684 Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman,
1685 S. J. (2017). The successor representation in human reinforcement learning. *Nature*
1686 *Human Behavior*, 1:680–692.
- 1687 Moran, R. and Goshen-Gottstein, Y. (2014). The conditional-recency dissociation is con-
1688 founded with nominal recency: should unitary models of memory still be devaluated?
1689 *Psychonomic Bulletin and Review*, 21:332–343.
- 1690 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental*
1691 *Psychology: General*, 64:482–488.
- 1692 Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy
1693 of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1694 Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhi-
1695 bitionless spreading activation and limited-capacity attention. *Journal of Experimental*
1696 *Psychology: General*, 106(3):226–254.
- 1697 Oberauer, K. and Lewandowsky, S. (2008). Forgetting in immediate serial recall: decay,
1698 temporal distinctiveness, or interference? *Psychological Review*, 115(3):544–576.
- 1699 Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., and Radvansky, G. A.
1700 (2016). Event boundaries and memory improvement. *Cognition*, 148:136–144.
- 1701 Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of
1702 context. *Trends in Cognitive Sciences*, 12:24–30.
- 1703 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). Task context and organization in
1704 free recall. *Neuropsychologia*, 47:2158–2163.

- 1705 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly*
1706 *Journal of Experimental Psychology*, 17:132–138.
- 1707 Puff, C. R. (1974). A consolidated theoretical view of stimulus-list organization effects in
1708 free recall. *Psychological Reports*, 34:275–288.
- 1709 Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of
1710 associative memory. In Bower, G. H., editor, *The Psychology of Learning and Motivation:*
1711 *Advances in Research and Theory*, volume 14, pages 207–262. Academic Press, New York,
1712 NY.
- 1713 Rabinowitz, J. C. (1986). Priming in episodic memory. *Journal of Gerontology*, 41:204–213.
- 1714 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:
1715 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.
- 1716 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition.
1717 *Current Opinion in Behavioral Sciences*, 17:133–140.
- 1718 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior.
1719 *Nature Reviews Neuroscience*, 13:713–726.
- 1720 Reinitz, M. T., Lammers, W. J., and Cochran, B. P. (1992). Memory-conjunction errors:
1721 miscombination of stored stimulus features can produce illusions of memory. *Memory*
1722 *and Cognition*, 20:1–11.
- 1723 Rissman, J., Eliassen, J. C., and Blumstein, S. E. (2003). An event-related fMRI investigation
1724 of implicit semantic priming. *Journal of Cognitive Neuroscience*, 15(8):1160–1175.
- 1725 Romney, A. K., Brewer, D. D., and Batchelder, W. H. (1993). Predicting clustering from
1726 semantic structure. *Psychological Science*, 4:28–34.

- 1727 Sahakyan, L. and Kelley, C. M. (2002). A contextual change account of the directed
1728 forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
1729 28(6):1064–1072.
- 1730 Sahakyan, L. and Smith, J. R. (2014). A long time ago, in a context far, far away: Retro-
1731 spective time estimates and internal context change. *Journal of Experimental Psychology:*
1732 *Learning, Memory, and Cognition*, 40(1):86–93.
- 1733 Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclo-*
1734 *pedic Reference*, 3:501–506.
- 1735 Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008). A context-based theory of
1736 recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.
- 1737 Sederberg, P. B., Miller, J. F., Howard, W. H., and Kahana, M. J. (2010). The tempo-
1738 ral contiguity effect predicts episodic memory performance. *Memory and Cognition*,
1739 38(6):689–699.
- 1740 Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of
1741 time. *Neural Computation*, 24:134–193.
- 1742 Shapiro, S. I. (1970). Isolation effects, free recall, and organization. *Journal of Psychology*,
1743 24:178–183.
- 1744 Sirotin, Y. B., Kimball, D. R., and Kahana, M. J. (2005). Going beyond a single list: modeling
1745 the effects of prior experience on episodic free recall. *Psychonomic Bulletin and Review*,
1746 12(5):787–805.
- 1747 Slamecka, N. J. and Barlow, W. (1979). The role of semantic and surface features in word
1748 repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 18:617–627.

- 1749 Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: a review and
1750 meta-analysis. *Psychonomic Bulletin and Review*, 8(2):203–220.
- 1751 Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009). A
1752 Bayesian analysis of dynamics in free recall. *Advances in Neural Information Processing*
1753 *Systems*, 22.
- 1754 Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., and Zacks, J. M. (2011).
1755 Changes in events alter how people remember recent information. *Journal of Cognitive*
1756 *Neuroscience*, 23(5):1052–1064.
- 1757 Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception
1758 affect memory encoding and updating. *Journal of Experimental Psychology: General*,
1759 138(2):236–257.
- 1760 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in*
1761 *Cognitive Sciences*, 22(3):201–212.
- 1762 Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *The*
1763 *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37:571–
1764 590.
- 1765 Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P.,
1766 and Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, 316(5821):76–
1767 82.
- 1768 Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press, New York, NY.
- 1769 Tulving, E. and Schacter, D. L. (1991). Priming and human memory systems. *Science*,
1770 247:301–305.

- 1771 Watkins, P. C., Mathews, A., Williamson, D. A., and Fuller, R. D. (1992). Mood-congruent
1772 memory in depression: emotional priming or elaboration? *Journal of Abnormal Psychol-*
1773 *ogy*, 101(3):581–586.
- 1774 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American*
1775 *Journal of Psychology*, 35:396–401.
- 1776 Whitely, P. L. (1927). The dependence of learning and recall upon prior intellectual activi-
1777 ties. *Journal of Experimental Psychology: General*, 10:489–508.
- 1778 Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming.
1779 *Current Opinion in Neurobiology*, 8(2):227–233.
- 1780 Xu, X., Zhu, Z., and Manning, J. R. (2023). The psychological arrow of time drives
1781 temporal asymmetries in retrodicting versus predicting narrative events. *PsyArXiv*,
1782 page doi.org/10.31234/osf.io/yp2qu.
- 1783 Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., and Hasson, U.
1784 (2017). Same story, different story: the neural representation of interpretive frameworks.
1785 *Psychological Science*, 28(3):307–319.
- 1786 Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., and
1787 Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal*
1788 *of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792–804.
- 1789 Zhang, Q., Griffiths, T. L., and Norman, K. A. (2023). Optimal policies for free recall.
1790 *Psychological Review*, 130(4):1104–1125.
- 1791 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018).
1792 Is automatic speech-to-text transcription ready for use in psychological experiments?
1793 *Behavior Research Methods*, 50:2597–2605.

- 1794 Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation
1795 models in narrative comprehension: an event-indexing model. *Psychological Science*,
1796 6(5):292–297.
- 1797 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension
1798 and memory. *Psychological Bulletin*, 123(2):162–185.