

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

3 Jeremy R. Manning<sup>1,\*</sup>, Emily C. Whitaker<sup>1</sup>, Paxton C. Fitzpatrick<sup>1</sup>,  
Madeline R. Lee<sup>1</sup>, Allison M. Frantz<sup>1</sup>, Bryan J. Bollinger<sup>1</sup>,  
Darya Romanova<sup>1</sup>, Campbell E. Field<sup>1</sup>, and Andrew C. Heusser<sup>1,2</sup>

<sup>1</sup>Dartmouth College

<sup>2</sup>Akili Interactive

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

4 **Abstract**

5 We perceive, interpret, and remember 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 is that our moment-by-  
22 moment subjective experiences are shaped in part by the idiosyncratic prior experiences,  
23 memories, goals, thoughts, expectations, and emotions that we bring with us into the  
24 present moment. These factors collectively define a *context* for our experiences (Manning,  
25 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  
36 interact 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, participants study  
42 lists of items and are instructed to recall the items in any order they choose. The orders  
43 in which words come to mind can provide insights into how participants have organized

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

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

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

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

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

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

116 From a theoretical perspective, we are interested in several core questions organized  
117 around the central theme of how structure in our experiences affect how we remember  
118 *those* experiences, and also how we remember *future* experiences (which may or may not

119 exhibit similar structure). For example, when we distill participants' experiences down  
120 to simple word lists that vary (meaningfully) along just a few feature dimensions, are  
121 there important differences in which dimensions influence participants' memories? Or  
122 are all features essentially "equally" influential? Further, are there differences in how  
123 specific features influence participants' memories for ongoing versus future experiences?  
124 Are there interaction effects between different features, or do people appear to treat each  
125 feature independently? And are there individual differences in how people organize their  
126 memories, or in how people are influenced by our experimental manipulations? If so,  
127 what are those differences and which aspects of memory do they affect?

## 128 **Materials and methods**

### 129 **Participants**

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

136 Participants either received course credit or a one-time \$10 payment for enrolling in  
137 our study. We asked each participant to fill out a demographic survey that included  
138 questions about their age, gender, ethnicity, race, education, vision, reading impairments,  
139 medications or recent injuries, coffee consumption on the day of testing, and level of  
140 alertness at the time of testing. All components of the demographics survey were optional.  
141 One participant elected not to fill out any part of the demographic survey, and all other

142 participants answered some or all of the survey questions.

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

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

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

182 We dropped from our dataset the one participant who reported having abnormal color  
183 vision, as well as 38 participants whose data were corrupted due to technical failures while  
184 running the experiment or during the daily database merges. In total, this left usable data  
185 from 452 participants, broken down by experimental condition as follows: feature rich (67  
186 participants), reduced (61 participants), reduced (early) (42 participants), reduced (late)  
187 (41 participants), category (30 participants), size (30 participants), length (30 participants),  
188 first letter (30 participants), color (31 participants), location (30 participants), and adaptive  
189 (60 participants). The participant who declined to fill out their demographic survey  
190 participated in the location condition, and we verified verbally that they had normal color  
191 vision and no significant reading impairments.



## 192 **Experimental design**

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

## 207 **Stimuli**

208 The stimuli in our paradigm were 256 English words selected in a previous study (Ziman  
209 et al., 2018). The words all referred to concrete nouns, and were chosen from 15 unique se-  
210 mantic categories: body parts, building-related, cities, clothing, countries, flowers, fruits,  
211 insects, instruments, kitchen-related, mammals, (US) states, tools, trees, and vegetables.  
212 We also tagged each word according to the approximate size of the object the word referred  
213 to. Words were labeled as “small” if the corresponding object was likely able to “fit in  
214 a standard shoebox” or “large” if the object was larger than a shoebox. Most semantic  
215 categories comprised words that reflected both “small” and “large” object sizes, but sev-



**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.

216 eral included only one or the other (e.g., all countries, US states, and cities are larger than  
217 a shoebox; mean number of different sizes per category: 1.33; standard deviation: 0.49).  
218 The numbers of words in each semantic category also varied from 12–28 (mean number of  
219 words per category: 17.07; standard deviation number of words: 4.65). We also identified  
220 lexicographic features for each word, including the words’ first letters and lengths (i.e.,  
221 number of letters). Across all categories, all possible first letters were represented except  
222 for ‘Q’ (average number of unique first letters per category: 11; standard deviation: 2  
223 letters). Word lengths ranged from 3–12 letters (average: 6.17 letters; standard deviation:  
224 2.06 letters).

225 We assigned the categorized words into a total of 16 lists with several constraints. First,  
226 we required that each list contained words from exactly four unique categories, each with  
227 exactly four exemplars from each category. Second, we required that (across all words  
228 on the list) at least one instance of both object sizes were represented. On average, each  
229 category was represented in 4.27 lists (standard deviation: 1.16 lists). Aside from these  
230 two constraints, we assigned each word to a unique list. After random assignment, each  
231 list contained words with an average of 11.13 unique starting letters (standard deviation:  
232 1.15 letters) and an average word length of 6.17 letters (standard deviation: 0.34 letters).

233 The above assignments of words to lists was performed once across all participants,  
234 such that every participant studied the same set of 16 lists. In every condition we random-  
235 ized the study order of these lists across participants. For participants in most conditions,  
236 on some or all of the lists, we also randomly varied two additional visual features associ-  
237 ated with each word: the presentation font color, and the word’s onscreen location. These  
238 attributes were assigned independently for each word (and for every participant). These  
239 visual features were varied for words in all lists and conditions except for the “reduced”  
240 condition (all lists), the first eight lists of the “reduced (early)” condition, and the last eight

241 lists of the “reduced (late)” condition. In these latter cases, words were all presented in  
242 black at the center of the experimental computer’s display.

243 To select a random font color for each word, we drew three integers uniformly and  
244 at random from the interval  $[0, 255]$ , corresponding to the red (r), green (g), and blue  
245 (b) color channels for that word. To assign random presentation locations to each word,  
246 we selected two floating point numbers uniformly and at random (one for the word’s  
247 horizontal  $x$ -coordinate and the other for its vertical  $y$ -coordinate). The bounds of these  
248 coordinates were selected to cover the entire visible area of the display without cutting off  
249 any part of the words. The words were shown on 27-in (diagonal) Retina 5K iMac displays  
250 (resolution:  $5120 \times 2880$  pixels).

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

## 256 **Real-time speech-to-text processing**

257 Our experimental paradigm incorporates the Google Cloud Speech API speech-to-text en-  
258 gine (Halpern et al., 2016) to automatically transcribe participants’ verbal recalls into text.  
259 This allows recalls to be transcribed in real time—a distinguishing feature of the experi-  
260 ment; in typical verbal recall experiments, the audio data must be parsed and transcribed  
261 manually. In prior work, we used a similar experimental setup (equivalent to the “re-  
262 duced” condition in the present study) to verify that the automatically transcribed recalls  
263 were sufficiently close to human-transcribed recalls to yield reliable data (Ziman et al.,  
264 2018). This real-time speech processing component of the paradigm plays an important

265 role in the “adaptive” condition of the experiment, as described below.

#### 266 **Random conditions (Fig. 1, top four rows)**

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

275 We also designed two conditions where we varied the words’ visual appearances across  
276 lists. In the *reduced (early)* condition, we followed the “reduced” procedure (presenting  
277 each word in black at the center of the screen) for early lists, and followed the “feature rich”  
278 procedure (presenting each word in a random color and location) for late lists. Finally, in  
279 the *reduced (late)* condition, we followed the feature rich procedure for early lists and the  
280 reduced procedure for late lists.

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

282 Each of six *order manipulation* conditions used a different feature-based sorting procedure  
283 to order words on early lists, where each sorting procedure relied on one relevant feature  
284 dimension. All of the irrelevant features varied freely across words on early lists, in that  
285 we did not consider irrelevant features in ordering the early lists. However, we note that  
286 some features were correlated—for example, some semantic categories of words referred  
287 to objects that tended to be a particular size, which meant that category and size were not

288 fully independent. On late lists, the words were always presented in a randomized order  
289 (chosen anew for each participant). In all of the order manipulation conditions, we varied  
290 words' font colors and onscreen locations, as in the feature rich condition.

291 **Defining feature-based distances.** Sorting words according to a given relevant feature  
292 requires first defining a distance function for quantifying the dissimilarity between each  
293 pair of features. This function varied according to the type of feature under consideration.  
294 Semantic features (category and size) are *categorical*. For these features, we defined a  
295 binary distance function: two words were considered to “match” (i.e., have a distance of  
296 0) if their labels were the same (i.e., both from the same semantic category or both of the  
297 same size). If two words' labels were different for a given feature, we defined the words  
298 to have a distance of 1 for that feature. Lexicographic features (length and first letter)  
299 are *discrete*. For these features we defined a discrete distance function. Specifically, we  
300 defined the distance between two words as either the absolute difference between their  
301 lengths, or the absolute distance between their starting letters in the English alphabet,  
302 respectively. For example, two words that started with the same letter would have a “first  
303 letter” distance of 0, and a pair of words starting with ‘J’ and ‘A’ would have a first letter  
304 distance of 9. Because words' lengths and letters' positions in the alphabet are always  
305 integers, these discrete distances always take on integer values. Finally, the visual features  
306 (color and location) are *continuous* and *multivariate*, in that each “feature” is defined by  
307 multiple (positive) real values. We defined the “color” and “location” distances between  
308 two words as the Euclidean distances between their  $(r, g, b)$  color or  $(x, y)$  location vectors  
309 (specified in inches), respectively. Therefore, the color and location distance measures  
310 always take on non-negative real values (upper-bounded at 441.67 for color, or 27 in for  
311 location, reflecting the distances between the corresponding maximally different vectors).

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

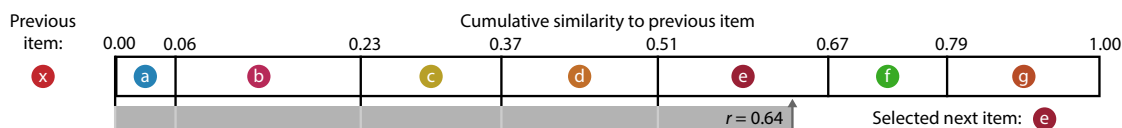
$$\text{similarity}(a, b) = \exp\{-\tau \cdot \text{distance}(a, b)\}, \quad (1)$$

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

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

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

326 As illustrated in Figure 2, we use these normalized similarity scores to construct a  
 327 sequence of “sticks” that we lay end to end in a line. Each of the  $n$  sticks corresponds to a  
 328 single to-be-presented word, and the stick lengths are proportional to the relative similar-  
 329 ities between each word’s feature value(s) and the feature value(s) of the just-presented  
 330 word. We choose the next to-be-presented word by moving an indicator along the set of  
 331 sticks, by a distance chosen uniformly and at random on the interval  $[0, 1]$ . We select the



**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).

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 example “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 lists, in a randomized order. We varied the words’ colors and locations for every word presentation, as in the feature rich and order manipulation conditions.

All participants in the adaptive condition began the experiment by studying a set of four *initialization* lists. Words and features on these lists were presented in a randomized order (computed independently for each participant). These initialization lists were used to estimate each participant’s “memory fingerprint,” defined below. At a high level,



347 a participant’s memory fingerprint describes how they prioritize or consider different  
348 semantic, lexicographic, and/or visual features when they organize their memories.

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

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

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

372 given feature dimension. We repeated this process for each feature dimension in turn to  
373 obtain a single uncorrected clustering score for each list, for each feature dimension.

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

384 **Permutation-corrected feature clustering scores.** Suppose that two lists contain unequal  
385 numbers of items of each size. For example, suppose that list *A* contains all "large" items,  
386 whereas list *B* contains an equal mix of "large" and "small" items. For a participant  
387 recalling list *A*, any correctly recalled item will necessarily match the size of the previous  
388 correctly recalled item. In other words, successively recalling several list *A* items of the  
389 same size is essentially meaningless, since *any* correctly recalled list *A* word will be large.  
390 In contrast, successively recalling several list *B* items of the same size *could* be meaningful,  
391 since (early in the recall sequence) the yet-to-be-recalled items come from a mix of sizes.  
392 However, once all of the small items on list *B* have been recalled, the best possible next  
393 matching recall will be a large item. All subsequent correct recalls must also be large  
394 items—so for those later recalls it becomes difficult to determine whether the participant  
395 is successively recalling large items because they are organizing their memories according

396 to size, or (alternatively), whether they are simply recalling the yet-to-be-recalled items  
397 in a random order. In general, the precise order and blend of feature values expressed  
398 in a given list, the order and number of correct recalls a participant makes, the number  
399 of intervening presentation positions between successive recalls, and so on, can all affect  
400 the range of clustering scores that are possible to observe for a given list. An uncorrected  
401 clustering score therefore conflates participants’ actual memory organization with other  
402 “nuisance” factors.

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

418 **Memory fingerprints.** We define each participant’s *memory fingerprint* as the set of their  
419 permutation-corrected clustering scores across all dimensions we tracked in our study,  
420 including their six feature-based clustering scores (category, size, length, first letter, color,

421 and location) and their temporal clustering score. Conceptually, a participant’s memory  
422 fingerprint describes their tendency to order in their recall sequences (and, presumably,  
423 organize in memory) the studied words along each dimension. To obtain stable estimates  
424 of these fingerprints for each participant, we averaged their clustering scores across lists.  
425 We also tracked and characterized how participants’ fingerprints changed across lists (e.g.,  
426 Figs. 6, S8).

427 **Online “fingerprint” analysis.** The presentation orders of some lists in the adaptive  
428 condition of our experiment (see *Adaptive condition*) were sorted according to participants’  
429 *current* memory fingerprint, estimated using all of the lists they had studied up to that point  
430 in the experiment. Because our experiment incorporated a speech-to-text component, all  
431 of the behavioral data for each participant could be analyzed just a few seconds after the  
432 conclusion of the recall intervals for each list. We used the Quail Python package (Heusser  
433 et al., 2017) to apply speech-to-text algorithms to the just-collected audio data, aggregate  
434 the data for the given participant, and estimate the participant’s memory fingerprint  
435 using all of their available data up to that point in the experiment. Two aspects of our  
436 implementation are worth noting. First, because memory fingerprints are computed  
437 independently for each list and then averaged across lists, the already-computed memory  
438 fingerprints for earlier lists could be cached and loaded as needed in future computations.  
439 This meant that our computations pertaining to updating our estimate of a participant’s  
440 memory fingerprint only needed to consider data from the most recent list. Second, each  
441 element of the null distributions of uncorrected fingerprint scores (see *Permutation-corrected*  
442 *feature clustering scores*) could be estimated independently from the others. This enabled  
443 us to make use of the testing computers’ multi-core CPU architectures by considering (in  
444 parallel) elements of the null distributions in batches of eight (i.e., the number of CPU  
445 cores on each testing computer). Taken together, we were able to compress the relevant

446 computations into just a few seconds of computing time. The combined processing time for  
447 the speech-to-text algorithm, fingerprint computations, and permutation-based ordering  
448 procedure (described next) easily fit within the inter-list intervals, where participants  
449 paused for a self-paced break before moving on to study and recall the next list.

450 **Ordering “stabilize” and “destabilize” lists by an estimated fingerprint.** In the adap-  
451 tive condition of our experiment, the presentation orders for *stabilize* and *destabilize* lists  
452 were chosen to either maximally or minimally (respectively) comport with participants’  
453 memory fingerprints. Given a participant’s memory fingerprint and a to-be-presented set  
454 of items, we designed a permutation-based procedure for ordering the items. First, we  
455 dropped from the participant’s fingerprint the temporal clustering score. For the remain-  
456 ing feature dimensions, we arranged the clustering scores in the fingerprint into a template  
457 vector,  $f$ . Second, we computed  $n = 2500$  random permutations of the to-be-presented  
458 items. These permutations served as candidate presentation orders. We sought to select  
459 the specific order that most (or least) closely matched  $f$ . Third, for each random permu-  
460 tation, we computed the (permutation-corrected) “fingerprint,” treating the permutation  
461 as though it were a potential “perfect” recall sequence. (We did not include temporal  
462 clustering scores in these fingerprints, since the temporal clustering score for every per-  
463 mutation is always equal to 1.) This yielded a “simulated fingerprint” vector,  $\hat{f}_p$  for each  
464 permutation  $p$ . We used these simulated fingerprints to select a specific permutation,  $i$ ,  
465 that either maximized (for stabilize lists) or minimized (for destabilize lists) the correlation  
466 between  $\hat{f}_i$  and  $f$ .

#### 467 **Computing low-dimensional embeddings of memory fingerprints**

468 Following some of our prior work (Heusser et al., 2021, 2018; Manning et al., 2022),  
469 we use low-dimensional embeddings to help visualize how participants’ memory fin-

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

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

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

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

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

## 515 **Analyses**

### 516 **Probability of $n^{\text{th}}$ recall curves**

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

532 We note that several other studies have used a slightly different approach to compute  
533 these curves, by correcting for the “availability” of a given word to be recalled. For  
534 example, if a participant recalls item 1, then item 2 on a given list, our approach places a  
535 0 into the item 1 column for that list when computing the “probability of second recall”  
536 curve. However, accounting for the fact that the participant had already recalled item  
537 1, an alternative approach (e.g., Farrell, 2010) would be to count the item 1 column as  
538 “unobserved” (i.e., missing data). Ultimately we chose to use the simpler variant of this  
539 approach in our work, but we direct the reader to further discussion of this issue in other



540 work (Farrell, 2014; Moran and Goshen-Gottstein, 2014).

### 541 **Lag-conditional response probability curve**

542 The lag-conditional response probability (lag-CRP) curve (Kahana, 1996) reflects the prob-  
543 ability of recalling a given item after the just-recalled item, as a function of their relative  
544 encoding positions (lag). In other words, a lag of 1 indicates that a recalled item was  
545 presented immediately after the previously recalled item, and a lag of  $-3$  indicates that a  
546 recalled item came three items before the previously recalled item. For each recall tran-  
547 sition (following the first recall), we computed the lag between the just-recalled word's  
548 presentation position and the next-recalled word's presentation position. We computed  
549 the proportions of transitions (between successively recalled words) for each lag, normaliz-  
550 ing for the total numbers of possible transitions. In carrying out this analysis, we excluded  
551 all incorrect recalls and repetitions (i.e., recalling a word that had already appeared pre-  
552 viously in the current recall sequence). This yielded, for each list, a 1 by number-of-lags  
553 ( $-15$  to  $+15$ ; 30 lags in total, excluding lags of 0) array of conditional probabilities. We  
554 averaged these probabilities across lists to obtain a single lag-CRP for each participant.  
555 Because transitions at large absolute lags are rare, these curves are typically displayed  
556 using range restrictions (Kahana, 2012).

### 557 **Serial position curve**

558 Serial position curves (Murdock, 1962) reflect the proportion of participants who remember  
559 each item as a function of the items' serial positions during encoding. For each participant,  
560 we initialized a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then,  
561 for each correct recall, we identified the presentation position of the word and entered a  
562 1 into that position (row: list; column: presentation position) in the matrix. This resulted

563 in a matrix whose entries indicated whether or not the words presented at each position,  
564 on each list, were recalled by the participant (depending on whether the corresponding  
565 entries were set to 1 or 0). Finally, we averaged over the rows of the matrix to yield a  
566 1 by 16 array representing the proportion of words at each position that the participant  
567 remembered.

### 568 **Identifying event boundaries**

569 We used the distances between feature values for successively presented words (see *Defin-*  
570 *ing feature-based distances*) to estimate “event boundaries” where the feature values changed  
571 more than usual (DuBrow and Davachi, 2016; Ezzyat and Davachi, 2011; Manning et al.,  
572 2016; Radvansky and Copeland, 2006; Swallow et al., 2011, 2009). For each list, for each  
573 feature dimension, we computed the distribution of distances between the feature values  
574 for successively presented words. We defined event boundaries (e.g., Fig. 3B) as occurring  
575 between any successive pair of words whose distances along the given feature dimension  
576 were greater than one standard deviation above the mean for that list. Note that, because  
577 event boundaries are defined for each feature dimension, each individual list may contain  
578 several sets of event boundaries, each at different moments in the presentation sequence  
579 (depending on the feature dimension of interest).

### 580 **Data and code availability**

581 All of the data analyzed in this manuscript, along with all of the code for carrying out the  
582 analyses may be found at <https://github.com/ContextLab/FRFR-analyses>.

## 583 Results

584 While holding the set of words (and the assignments of words to lists) constant, we  
585 manipulated two aspects of participants' experiences of studying each list. We sought to  
586 understand the effects of these manipulations on participants' memories for the studied  
587 words. First, we added two additional sources of visual variation to the individual word  
588 presentations: font color and onscreen location. Importantly, these visual features were  
589 independent of the meaning or semantic content of the words (e.g., word category, size  
590 of the referent, etc.) and of the lexicographic properties of the words (e.g., word length,  
591 first letter, etc.). We wondered whether this additional word-independent information  
592 might facilitate recall (e.g., by providing new or richer potential ways of organizing or  
593 retrieving memories of the studied words; Davachi et al., 2003; Drewnowski and Murdock,  
594 1980; Hargreaves et al., 2012; Madan, 2021; Meinhardt et al., 2020; Slamecka and Barlow,  
595 1979; Socher et al., 2009) or impair recall (e.g., by distracting or confusing participants  
596 with irrelevant information Lange, 2005; Marsh et al., 2012, 2015; Reinitz et al., 1992).  
597 Second, we manipulated the orders in which words were studied (and how those orderings  
598 changed over time). We wondered whether presenting the same list of words with different  
599 appearances (e.g., by manipulating font size and onscreen location) or in different orders  
600 (e.g., sorted along one feature dimension versus another) might serve to influence how  
601 participants organized their memories of the words (e.g., Manning et al., 2015; Polyn and  
602 Kahana, 2008). We also wondered whether some order manipulations might be temporally  
603 "sticky" by influencing how *future* lists were remembered (e.g., Baddeley, 1968; Darley  
604 and Murdock, 1971; Lohnas et al., 2010; Sirotin et al., 2005; Whitely, 1927).

605 To obtain a clean preliminary estimate of the consequences on memory of randomly  
606 varying the font colors and locations of presented words (versus holding the font color  
607 fixed at black, and holding the display locations fixed at the center of the display) we

608 compared participants' performance on the *feature rich* and *reduced* experimental condi-  
 609 tions (see *Random conditions*, Fig. S1). In the feature rich condition the words' colors and  
 610 locations varied randomly across words, and in the reduced condition words were always  
 611 presented in black, at the center of the display. Aggregating across all lists for each partic-  
 612 ipant, we found no difference in recall accuracy (i.e., the proportions of correctly recalled  
 613 words) for feature rich versus reduced lists ( $t(126) = -0.290, p = 0.772$ , Cohen's  $d$  ( $d$ ) =  
 614  $-0.051$ , bootstrap estimated 95% confidence interval (CI) =  $[-2.387, 1.768]$ ). However,  
 615 participants in the feature rich condition clustered their recalls substantially more along  
 616 every dimension we examined (temporal clustering:  $t(126) = 10.632, p < 0.001, d =$   
 617  $1.882$ , CI =  $[7.786, 14.386]$ ; semantic category clustering:  $t(126) = 10.148, p < 0.001, d =$   
 618  $1.796$ , CI =  $[7.324, 13.778]$ ; size clustering:  $t(126) = 12.033, p < 0.001, d = 2.129$ , CI =  
 619  $[9.030, 15.918]$ ; word length clustering:  $t(126) = 10.720, p < 0.001, d = 1.897$ , CI =  $[7.442, 15.174]$ ;  
 620 first letter clustering:  $t(126) = 6.679, p < 0.001, d = 1.182$ , CI =  $[4.490, 9.611]$ ; see *Permutation-*  
 621 *corrected feature clustering scores* for more information about how we quantified each par-  
 622 ticipant's clustering tendencies.) Taken together, these comparisons suggest that adding  
 623 new features changes how participants organize their memories of studied words, even  
 624 when those new features are independent of the words themselves and even when the new  
 625 features vary randomly across words. We found no evidence that those additional unin-  
 626 formative features were distracting (in terms of their impact on memory performance),  
 627 but they did affect participants' recall dynamics (measured via their clustering scores).

628 We also wondered whether adding these incidental visual features to later lists (after  
 629 the participants had already studied impoverished lists), or removing the visual features  
 630 from later lists (after the participants had already studied visually diverse lists) might  
 631 affect memory performance. In other words, we sought to test for potential effects of  
 632 changing the "richness" of participants' experiences over time. All participants stud-

633 ied and recalled a total of 16 lists; we defined *early* lists as the first eight lists and *late*  
 634 lists as the last eight lists each participant encountered. To help interpret our results,  
 635 we compared participants' memories on early versus late lists in the above feature rich  
 636 and reduced conditions. Participants in both conditions remembered more words on  
 637 early versus late lists (feature rich:  $t(66) = 4.553, p < 0.001, d = 0.233, CI = [2.427, 7.262]$ ;  
 638 reduced:  $t(60) = 2.434, p = 0.018, d = 0.134, CI = [0.493, 4.910]$ ). Participants in the  
 639 feature rich (but not reduced) conditions exhibited more temporal clustering on early  
 640 versus late lists (feature rich:  $t(66) = 2.268, p = 0.027, d = 0.181, CI = [0.437, 4.425]$ ; re-  
 641 duced:  $t(60) = 0.986, p = 0.328, d = 0.061, CI = [-0.897, 3.348]$ ). And participants in  
 642 both conditions tended to exhibit more semantic clustering on early versus late lists  
 643 (feature rich, category:  $t(66) = 3.684, p < 0.001, d = 0.220, CI = [1.733, 5.732]$ ; feature  
 644 rich, size:  $t(66) = 1.629, p = 0.108, d = 0.100, CI = [-0.207, 3.905]$ ; reduced, category:  
 645  $t(60) = 2.755, p = 0.008, d = 0.177, CI = [0.761, 5.189]$ ; reduced, size:  $t(60) = 3.081, p =$   
 646  $0.003, d = 0.201, CI = [1.210, 5.326]$ ). Participants in the reduced (but not feature rich)  
 647 conditions tended to exhibit more lexicographic clustering on early versus late lists (fea-  
 648 ture rich, word length:  $t(66) = -0.100, p = 0.921, d = -0.010, CI = [-2.217, 1.899]$ ; feature  
 649 rich, first letter:  $t(66) = -0.412, p = 0.681, d = -0.045, CI = [-2.461, 1.645]$ ; reduced,  
 650 word length:  $t(60) = 3.762, p < 0.001, d = 0.261, CI = [1.604, 6.821]$ ; reduced, first letter:  
 651  $t(60) = 1.721, p = 0.090, d = 0.175, CI = [-0.138, 4.098]$ ). Taken together, these comparisons  
 652 suggest that even when the presence or absence of incidental visual features is stable  
 653 across lists, participants still exhibit some differences in their performance and memory  
 654 organization tendencies for early versus late lists.

655 With these differences in mind, we next compared participants' memories on early ver-  
 656 sus late lists for two additional experimental conditions (see *Random conditions*, Fig. S1).  
 657 In a *reduced (early)* condition, we held the visual features constant on early lists, but al-

658 lowed them to vary randomly on late lists. In a *reduced (late)* condition, we allowed  
 659 the visual features to vary randomly on early lists, but held them constant on late  
 660 lists. Given our above findings that (a) participants tended to exhibit stronger clus-  
 661 tering effects on feature rich (versus reduced) lists, and (b) participants tended to re-  
 662 member more words and exhibit stronger clustering effects on early (versus late) lists,  
 663 we expected these early versus late differences to be enhanced in the reduced (early)  
 664 condition and diminished in the reduced (late) condition. However, to our surprise,  
 665 participants in *neither* condition exhibited reliable early versus late differences in accu-  
 666 racy (reduced (early):  $t(41) = 1.499, p = 0.141, d = 0.098, CI = [-0.345, 3.579]$ ; reduced  
 667 (late):  $t(40) = 1.462, p = 0.152, d = 0.121, CI = [-0.376, 2.993]$ ), temporal clustering (re-  
 668 duced (early):  $t(41) = 0.857, p = 0.396, d = 0.068, CI = [-1.012, 2.896]$ ; reduced (late):  
 669  $t(40) = 1.244, p = 0.221, d = 0.128, CI = [-0.894, 3.088]$ ), nor feature-based clustering (re-  
 670 duced (early), category:  $t(41) = 0.707, p = 0.484, d = 0.068, CI = [-1.314, 2.830]$ ; reduced  
 671 (early), size:  $t(41) = 0.803, p = 0.427, d = 0.079, CI = [-1.142, 2.953]$ ; reduced (early),  
 672 length:  $t(41) = 0.461, p = 0.648, d = 0.060, CI = [-1.545, 2.462]$ ; reduced (early), first  
 673 letter:  $t(41) = 0.781, p = 0.439, d = 0.101, CI = [-1.039, 2.881]$ ; reduced (late), category:  
 674  $t(40) = -0.101, p = 0.920, d = -0.009, CI = [-2.307, 1.776]$ ; reduced (late), size:  $t(40) =$   
 675  $0.555, p = 0.582, d = 0.058, CI = [-1.444, 2.274]$ ; reduced (late), length:  $t(40) =$   
 676  $0.146, d = 0.126, CI = [-0.444, 3.743]$ ; reduced (late), first letter:  $t(40) = -0.143, p =$   
 677  $0.887, d = -0.017, CI = [-2.204, 1.830]$ ). We hypothesized that adding or removing the  
 678 variability in the visual features was acting as a sort of “event boundary” between early  
 679 and late lists (e.g., Clewett et al., 2019; Radvansky and Copeland, 2006; Radvansky and  
 680 Zacks, 2017). In prior work, we (and others) have found that memories formed just af-  
 681 ter event boundaries can be enhanced (e.g., due to less contextual interference between  
 682 pre- and post-boundary items; Flores et al., 2017; Gold et al., 2017; Manning et al., 2016;

683 Pettijohn et al., 2016).

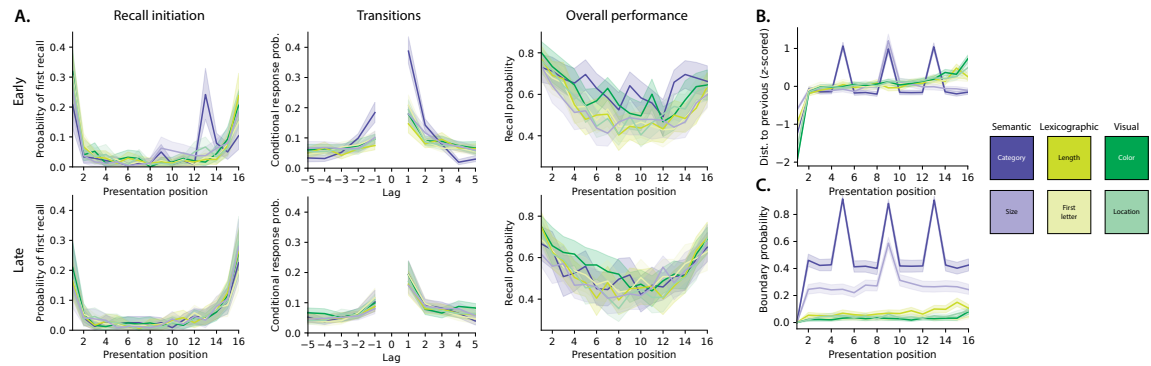
684 We found that *adding* incidental visual features on later lists that had not been present  
685 on early lists (as in the reduced (early) condition) served to enhance recall performance  
686 relative to conditions where all lists had the same blends of features (accuracy for feature  
687 rich versus reduced (early):  $t(107) = -2.230, p = 0.028, d = -0.439, CI = [-4.252, -0.229]$ ;  
688 reduced versus reduced (early):  $t(101) = -2.045, p = 0.043, d = -0.410, CI = [-3.826, 0.112]$ ;  
689 also see Fig. S3A). However, *subtracting* irrelevant visual features on later lists that *had*  
690 been present on early lists (as in the reduced (late) condition) did not appear to impact  
691 recall performance (accuracy for feature rich versus reduced (late):  $t(106) = -0.638, p =$   
692  $0.525, d = -0.126, CI = [-2.720, 1.362]$ ; reduced versus reduced (late):  $t(100) = -0.407, p =$   
693  $0.685, d = -0.082, CI = [-2.477, 1.626]$ ). These comparisons suggest that recall accuracy has  
694 a directional component: accuracy is affected differently by removing features later that  
695 had been present earlier versus adding features later that had *not* been present earlier. In  
696 contrast, we found that participants exhibited more temporal and feature-based clustering  
697 when we added incidental visual features to *any* lists (comparisons of clustering on feature  
698 rich versus reduced lists are reported above; temporal clustering in reduced versus reduced  
699 (early) and reduced versus reduced (late) conditions:  $ts \leq -9.885, ps < 0.001$ ; feature-based  
700 clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions:  
701  $ts \leq -4.555, ps < 0.001$ ). Temporal and feature-based clustering were not reliably different  
702 in the feature rich, reduced (early), and reduced (late) conditions (temporal clustering in  
703 feature rich versus reduced (early) and feature rich versus reduced (late) conditions:  $ts$   
704  $\geq -1.379, ps \geq 0.171$ ; feature-based clustering in feature rich versus reduced (early) and  
705 feature rich versus reduced (late) conditions:  $|ts| \leq 1.441, ps \geq 0.153$ ).

706 Taken together, our findings thus far suggest that adding item features that change  
707 over time, even when they vary randomly and independently of the items, can enhance

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

723 Across each of six order manipulation conditions, we sorted early lists by one feature  
 724 dimension but randomly ordered the items on late lists (see *Order manipulation conditions*;  
 725 features: category, size, length, first letter, color, and location). Participants in the category-  
 726 ordered condition showed an increase in memory performance on early lists (accuracy,  
 727 relative to early feature rich lists;  $t(95) = 3.034, p = 0.003, d = 0.667, CI = [1.048, 5.113]$ ).  
 728 Participants in the color-ordered condition also showed a trending increase in memory  
 729 performance on early lists (again, relative to early feature rich lists:  $t(96) = 1.850, p =$   
 730  $0.067, d = 0.402, CI = [-0.010, 3.712]$ ; Fig. 5A). Participants' performances on early lists in  
 731 all of the other order manipulation conditions were indistinguishable from performance  
 732 on the early feature rich lists ( $|t| \leq 1.013, ps \geq 0.314$ ). Participants in both of the semanti-

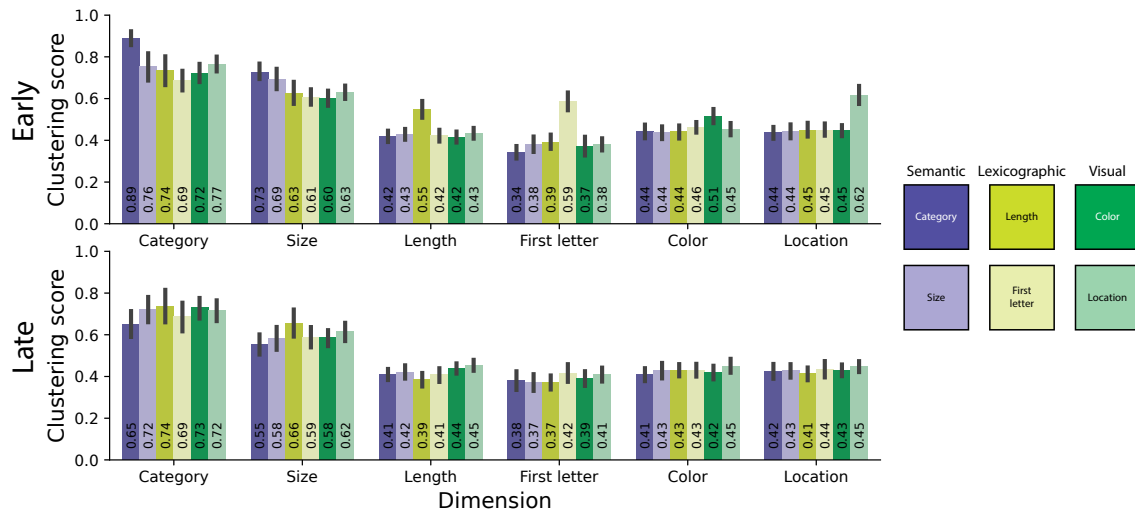




**Figure 3: Recall dynamics in feature rich free recall (order manipulation conditions).** **A.** Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling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

733 cally ordered conditions exhibited stronger temporal clustering on early lists (versus early  
 734 feature rich lists; category:  $t(95) = 8.813, p < 0.001, d = 1.936, CI = [6.793, 11.751]$ ; size:  
 735  $t(95) = 2.630, p = 0.010, d = 0.578, CI = [0.831, 4.866]$ ; Fig. 5B). Participants in the length-  
 736 ordered condition tended to exhibit *less* temporal clustering on early lists relative to early  
 737 feature rich lists ( $t(95) = -1.547, p = 0.125, d = -0.340, CI = [-3.693, 0.341]$ ), whereas  
 738 participants in the first letter-ordered condition exhibited stronger temporal clustering  
 739 on early lists ( $t(95) = 2.858, p = 0.005, d = 0.628, CI = [1.031, 4.886]$ ). Participants in the  
 740 visually ordered conditions exhibited more similar performance (accuracy) on early lists,  
 741 relative to early feature rich lists (we found a trending enhancement for participants in  
 742 the color-ordered condition:  $t(96) = 1.850, p = 0.067, d = 0.402, CI = [-0.010, 3.712]$ ; loca-  
 743 tion:  $t(95) = 0.043, p = 0.966, d = 0.010, CI = [-1.598, 1.729]$ ). Participants in the visually  
 744 ordered conditions also showed similar temporal clustering on early lists, relative to early  
 745 feature rich lists (color:  $t(96) = -1.339, p = 0.184, d = -0.291, CI = [-3.238, 0.394]$ , we found  
 746 a trending increase for participants in the location-ordered condition:  $t(95) = 1.705, p =$   
 747  $0.092, d = 0.374, CI = [-0.155, 3.521]$ ). We also compared feature-based clustering on early  
 748 lists across the order manipulation and feature rich conditions. Since these results were  
 749 similar across both semantic conditions (category and size), both lexicographic conditions  
 750 (length and first letter), and both visual conditions (color and location), here we aggre-  
 751 gate data from conditions that manipulated each of these three feature groupings in our  
 752 comparisons, to simplify the presentation. On early lists, participants in the semantically  
 753 ordered conditions exhibited stronger semantic clustering relative to participants in the  
 754 feature rich condition (category:  $t(125) = 2.722, p = 0.007, d = 0.484, CI = [0.827, 4.932]$ ;  
 755 size:  $t(125) = 3.866, p < 0.001, d = 0.687, CI = [2.020, 5.983]$ ), but showed no reliable dif-  
 756 ferences in lexicographic (length:  $t(125) = 0.521, p = 0.603, d = 0.093, CI = [-1.311, 2.333]$ ;  
 757 first letter:  $t(125) = -0.842, p = 0.401, d = -0.150, CI = [-2.825, 1.095]$ ) or visual (color:

$t(125) = -0.650, p = 0.517, d = -0.116, CI = [-2.680, 1.249]$ ; location:  $t(125) = -0.251, p =$   
 $0.802, d = -0.045, CI = [-2.257, 1.524]$ ) clustering. Similarly, participants in the lexico-  
graphically ordered conditions exhibited stronger (relative to feature rich participants)  
lexicographic clustering (length:  $t(125) = 3.682, p < 0.001, d = 0.655, CI = [1.890, 5.569]$ ;  
first letter:  $t(125) = 5.134, p < 0.001, d = 0.912, CI = [3.251, 7.258]$ ) on early lists, but showed  
no reliable differences in semantic (category:  $t(125) = -1.040, p = 0.301, d = -0.185, CI =$   
 $[-3.095, 1.092]$ ; size:  $t(125) = 0.006, p = 0.995, d = 0.001, CI = [-1.933, 1.952]$ ) or visual  
(color:  $t(125) = 0.092, p = 0.927, d = 0.016, CI = [-1.834, 1.867]$ ; location:  $t(125) = 0.407, p =$   
 $0.685, d = 0.072, CI = [-1.655, 2.463]$ ) clustering. And participants in the visually ordered  
conditions exhibited stronger visual clustering (again, relative to feature rich participants,  
and on early lists; color:  $t(126) = 2.022, p = 0.045, d = 0.358, CI = [0.056, 3.965]$ ; location:  
 $t(126) = 4.390, p < 0.001, d = 0.777, CI = [2.730, 6.199]$ ), but showed no reliable differ-  
ences in semantic (category:  $t(126) = 0.012, p = 0.991, d = 0.002, CI = [-1.988, 1.871]$ ;  
size:  $t(126) = -0.104, p = 0.917, d = -0.018, CI = [-2.166, 1.847]$ ) or lexicographic (length:  
 $t(126) = 0.592, p = 0.555, d = 0.105, CI = [-1.361, 2.420]$ ; first letter:  $t(126) = 0.040, p =$   
 $0.968, d = 0.007, CI = [-1.791, 1.863]$ ) clustering. Taken together, these order manipulation  
results suggest several broad patterns (Figs. 3A, 4). First, most of the order manipulations  
we carried out did *not* reliably affect overall recall performance. Second, most of the  
order manipulations increased participants' tendencies to temporally cluster their recalls.  
Third, all of the order manipulations enhanced participants' clustering of each condition's  
target feature (i.e., semantic manipulations enhanced semantic clustering, lexicographic  
manipulations enhanced lexicographic clustering, and visual manipulations enhanced vi-  
sual clustering; Fig. 5C) while leaving clustering along other feature dimensions roughly  
unchanged (i.e., semantic manipulations did not affect lexicographic or visual clustering,  
and so on). Although it is not possible to fully separate feature versus temporal clustering



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

when considering sorted lists, we used a permutation-based procedure to identify the degree of feature clustering over and above what could be accounted for by temporal clustering alone (see *Factoring out the effects of temporal clustering*). When we carried out this analysis (Fig. 5D), we found that participants exhibited more semantic clustering on semantically sorted lists than on randomly ordered lists, but the effects of the other order manipulations could not reliably be separated from temporal clustering alone (reliable comparisons are reported in the figure).

When we closely examined the sequences of words participants recalled from early order-manipulated lists (Fig. 3A, top panel), we noticed several differences from the dynamics of participants’ recalls of randomly ordered lists (Figs. S1, S7). One difference is that participants in the category condition (dark purple curves, Fig. 3) most often initiated recall with the fourth-from-last item (*Recall initiation*, top left panel), whereas participants

795 who recalled randomly ordered lists tended to initiate recall with either the first or last list  
796 items (Fig. S1, top left panel). We hypothesized that the participants might be “clumping”  
797 their recalls into groups of items that shared category labels. Indeed, when we com-  
798 pared the positions of feature changes in the study sequence (Fig. 3B; see *Identifying event*  
799 *boundaries*) with the positions of items participants recalled first, we noticed a striking  
800 correspondence in both semantic conditions. Specifically, on category-ordered lists, the  
801 category labels changed every four items on average (dark purple peaks in Fig. 3B), and  
802 participants also seemed to display an increased tendency (relative to other order manipu-  
803 lation and random conditions) to initiate recall of category-ordered lists with items whose  
804 study positions were integer multiples of four. Similarly, for size-ordered lists, the size la-  
805 bels changed every eight items on average (light purple peaks in Fig. 3B), and participants  
806 also seemed to display an increased tendency to initiate recall of size-ordered lists with  
807 items whose study positions were integer multiples of eight. A second striking difference  
808 is that participants in the category condition exhibited a much steeper lag-CRP (Fig. 3A,  
809 top middle panel) than participants in other conditions. (This is another expression of  
810 participants’ increased tendencies to temporally cluster their recalls on category-ordered  
811 lists, as we reported above.) Taken together, these order-specific idiosyncrasies suggest  
812 a hierarchical set of influences on participants’ memories. At longer timescales, “event  
813 boundaries” (to use the term loosely) can be induced across lists by adding or removing  
814 incidental visual features. At shorter timescales, “event boundaries” can be induced across  
815 items (within a single list) by adjusting how item features change throughout the list.

816 The above comparisons between memory performance on early lists in the order  
817 manipulation versus feature rich conditions highlight how sorted lists are remembered  
818 differently from random lists. We also wondered how sorting lists along each feature  
819 dimension influenced memory relative to sorting lists along the other feature dimen-

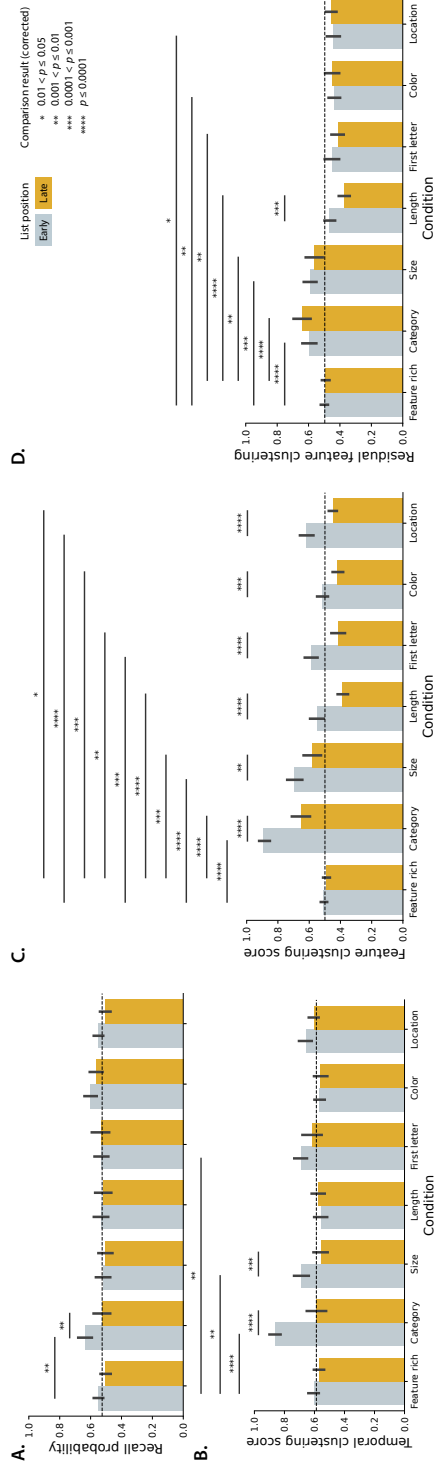
sions. Participants trended towards remembering early lists that were sorted semantically better than lexicographically sorted lists ( $t(118) = 1.936, p = 0.055, d = 0.353, CI = [0.057, 3.916]$ ). Participants also remembered visually sorted lists better than lexicographically sorted lists ( $t(119) = 2.145, p = 0.034, d = 0.390, CI = [0.208, 4.254]$ ). 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 either lexicographically ( $t(118) = 5.620, p < 0.001, d = 1.026, CI = [3.486, 8.010]$ ) or visually ( $t(119) = 6.613, p < 0.001, d = 1.202, CI = [4.481, 9.464]$ ) sorted lists, but did not show reliable differences in temporal clustering on lexicographically versus visually sorted lists ( $t(119) = 0.589, p = 0.557, d = 0.107, CI = [-1.336, 2.539]$ ). 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:  $t(118) = 3.972, p < 0.001, d = 0.738, CI = [2.145, 6.296]$ ) sorted lists; more lexicographic clustering on lexicographically sorted lists than semantically (length:  $t(118) = 3.390, p < 0.001, d = 0.619, CI = [1.499, 5.661]$ ; first letter:  $t(118) = 5.705, p < 0.001, d = 1.042, CI = [3.841, 7.790]$ ) or visually (length:  $t(119) = 3.399, p < 0.001, d = 0.618, CI = [1.500, 5.527]$ ; first letter:  $t(119) = 4.859, p < 0.001, d = 0.883, CI = [2.860, 6.849]$ ) sorted lists; and more visual clustering on visually sorted lists than semantically (color:  $t(119) = 2.673, p = 0.009, d = 0.486, CI = [0.848, 4.567]$ ; location:  $t(119) = 4.499, p < 0.001, d = 0.818, CI = [2.721, 6.399]$ ) or lexicographically (color:  $t(119) = 1.988, p = 0.049, d = 0.361, CI = [0.102, 3.894]$ ; location:  $t(119) = 3.966, p < 0.001, d = 0.721, CI = [2.099, 5.862]$ ) sorted lists. In summary, sorting lists by different features appeared to have slightly different effects on overall memory performance and temporal clustering. Participants also tended to cluster their recalls along a given fea-

ture dimension more when the studied lists were (versus were not) sorted along that dimension.

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

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

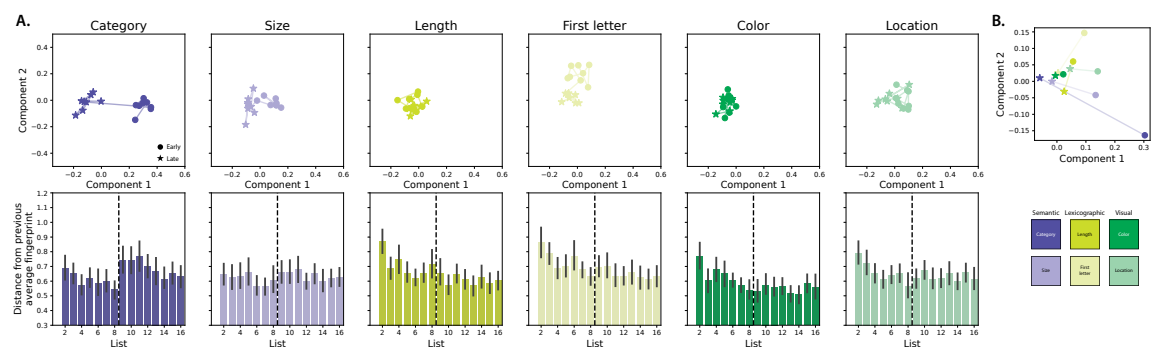


**Figure 5: Recall probability and clustering scores on early and late lists.** The bar heights display the average (across participants) recall probabilities (A.), temporal clustering scores (B.), feature clustering scores (C.), and residual feature clustering scores (after factoring out temporal clustering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



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

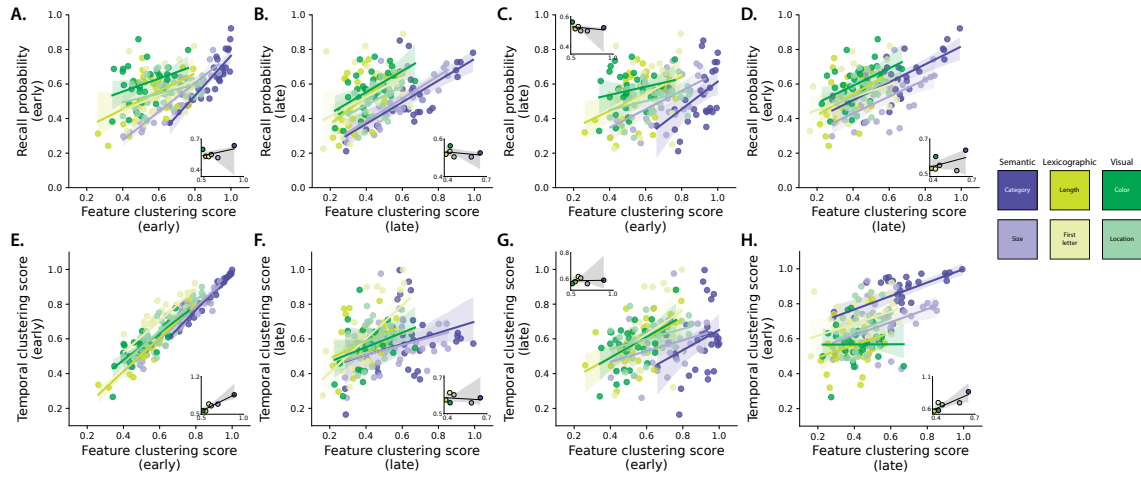
When we examined the data at the level of individual participants (Figs. 7 and 8), a



**Figure 6: Memory fingerprint dynamics (order manipulation conditions).** **A.** Each column (and color) reflects an experimental condition. In the top panels, each marker displays a 2D projection of the (across-participant) average memory fingerprint for one list. Order manipulation (early) lists are denoted by circles and randomly ordered (late) lists are denoted by stars. All of the fingerprints (across all conditions and lists) are projected into a common space. The bar plots in the bottom panels display the Euclidean distances of the per-list memory fingerprints to the list 0 fingerprint, for each condition. Error bars denote bootstrap-estimated 95% confidence intervals. The dotted vertical lines denote the boundaries between early and late lists. **B.** In this panel, the fingerprints for early (circle) and late (star) lists are averaged across lists and participants before projecting the fingerprints into a (new) 2D space. See Figure S8 for analogous plots for the random conditions.

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 individually 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 conditions 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 performance (Fig. 7C, D). Participants who showed stronger feature clustering on early lists tended to recall more

908 items on late lists (across conditions:  $r(179) = 0.492, p < 0.001$ ; all conditions individually:  
 909  $r_s \geq 0.462$ , all  $p_s \leq 0.010$ ). Participants who recalled more items on early lists also tended  
 910 to show stronger feature clustering on late lists (across conditions:  $r(179) = 0.280, p <$   
 911  $0.001$ ; all non-visual conditions:  $r_s \geq 0.445$ , all  $p_s \leq 0.014$ ; color:  $r(29) = 0.298, p =$   
 912  $0.103$ ; location:  $r(28) = 0.354, p = 0.055$ ). Neither of these effects showed condition-level  
 913 trends (early feature clustering versus late recall probability:  $r(4) = -0.299, p = 0.565$ ;  
 914 early recall probability versus late feature clustering:  $r(4) = 0.400, p = 0.432$ ). We also  
 915 looked for associations between feature clustering and temporal clustering. Across every  
 916 order manipulation condition, participants who exhibited stronger feature clustering also  
 917 exhibited stronger temporal clustering. For early lists (Fig. 7E), this trend held overall  
 918 ( $r(179) = 0.924, p < 0.001$ ), for each condition individually (all  $r_s \geq 0.822$ , all  $p_s < 0.001$ ),  
 919 and across conditions ( $r(4) = 0.964, p = 0.002$ ). For late lists (Fig. 7F), the results were  
 920 more variable (overall:  $r(179) = 0.348, p < 0.001$ ; all non-visual conditions:  $r_s \geq 0.382$ ,  
 921 all  $p_s \leq 0.037$ ; color:  $r(29) = 0.453, p = 0.011$ ; location:  $r(28) = 0.190, p = 0.314$ ; across-  
 922 conditions:  $r(4) = -0.036, p = 0.945$ ). While less robust than the carryover associations  
 923 between feature clustering and recall performance, we also observed some carryover  
 924 associations between feature clustering and temporal clustering (Fig. 7G, H). Participants  
 925 who showed stronger feature clustering on early lists trended towards showing stronger  
 926 temporal clustering on later lists (overall:  $r(179) = 0.464, p < 0.001, CI = [0.321, 0.582]$ ; for  
 927 individual conditions: all  $r_s \geq 0.377$ , all  $p_s \leq 0.040$ ; across conditions:  $r(4) = 0.451, p =$   
 928  $0.369, CI = [-0.986, 0.998]$ ). And participants who showed stronger temporal clustering  
 929 on early lists trended towards showing stronger feature clustering on later lists (overall:  
 930  $r(179) = 0.266, p < 0.001, CI = [0.129, 0.396]$ ; for individual conditions: all  $r_s \geq 0.298$ , all  
 931  $p_s \leq 0.110$ ; across conditions:  $r(4) = 0.064, p = 0.903, CI = [-0.972, .]$ ). Taken together,  
 932 the results displayed in Figure 7 show that participants who were more sensitive to the



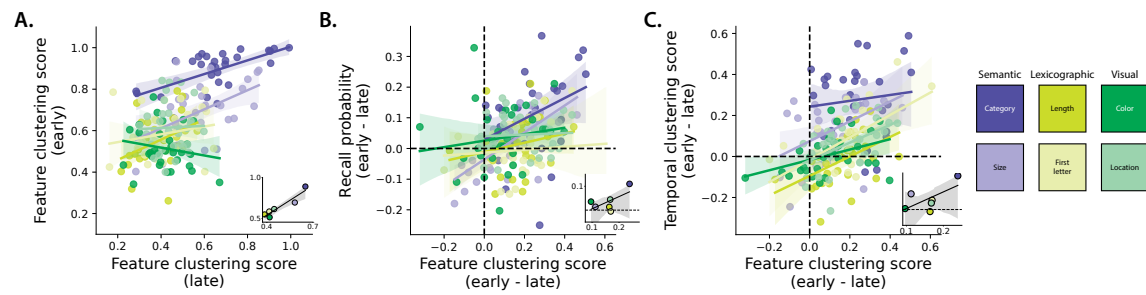
**Figure 7: Interactions between feature clustering, recall probability, and contiguity.** **A.** Recall probability versus feature clustering scores for order manipulation (early) lists. **B.** Recall probability versus feature clustering for randomly ordered (late) lists. **C.** Recall probability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

933 order manipulations (i.e., participants who showed stronger feature clustering for their  
 934 condition's feature on early lists) remembered more words and showed stronger temporal  
 935 clustering. These associations also appeared to carry over across lists, even when the items  
 936 on later lists were presented in a random order.

937 If participants show different sensitivities to order manipulations, how do their be-  
 938 haviors carry over to later lists? We found that participants who showed strong feature  
 939 clustering on early lists often tended to show strong feature clustering on late lists (Fig. 8A;  
 940 overall across participants and conditions:  $r(179) = 0.591, p < 0.001, CI = [0.472, 0.682]$ ;  
 941 category:  $r(28) = 0.590, p < 0.001, CI = [0.354, 0.756]$ ; size:  $r(28) = 0.488, p = 0.006, CI =$

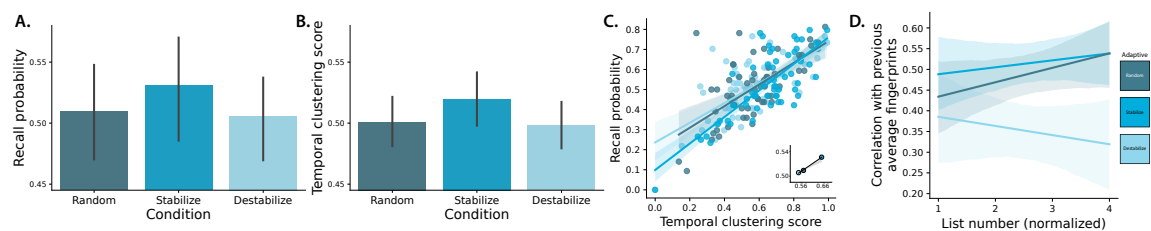
942 [0.134, 0.732]; length:  $r(28) = 0.384, p = 0.036, CI = [0.040, 0.681]$ ; first letter:  $r(28) =$   
 943  $0.202, p = 0.284, CI = [-0.273, 0.620]$ ; color:  $r(29) = -0.183, p = 0.325, CI = [-0.562, 0.258]$ ;  
 944 location:  $r(28) = 0.031, p = 0.870, CI = [-0.240, 0.296]$ ; across conditions:  $r(4) = 0.942, p =$   
 945  $0.005, CI = [0.442, 1.000]$ ). Although participants tended to show weaker feature clustering  
 946 on late lists (Fig. 6) on *average*, the associations between early and late lists for individual  
 947 participants suggests that some influence of early order manipulations may linger on late  
 948 lists. We found that participants who exhibited larger carryover in feature clustering (i.e.,  
 949 continued to show strong feature clustering on late lists) for the semantic order manip-  
 950 ulations (but not other manipulations) also tended to show a smaller decrease in recall  
 951 on early versus late lists (Fig. 8B; overall:  $r(179) = 0.307, p < 0.001, CI = [0.148, 0.469]$ ;  
 952 category:  $r(28) = 0.350, p = 0.058, CI = [0.050, 0.642]$ ; size:  $r(28) = 0.708, p < 0.001, CI =$   
 953  $[0.472, 0.862]$ ; length:  $r(28) = 0.205, p = 0.276, CI = [-0.109, 0.492]$ ; first letter:  $r(28) =$   
 954  $0.081, p = 0.672, CI = [-0.433, 0.597]$ ; color:  $r(29) = 0.155, p = 0.406, CI = [-0.174, 0.541]$ ;  
 955 location:  $r(28) = 0.052, p = 0.787, CI = [-0.307, 0.360]$ ; across conditions:  $r(4) = 0.635, p =$   
 956  $0.176, CI = [-0.924, 0.981]$ . Participants who exhibited larger carryover in feature cluster-  
 957 ing also tended to show stronger temporal clustering on late lists (relative to early lists)  
 958 for all but the category condition (Fig. 8C; overall:  $r(179) = 0.434, p < 0.001$ ; category:  
 959  $r(28) = 0.229, p = 0.223$ ; all non-category conditions: all  $rs \geq 0.448$ , all  $ps \leq 0.012$ ; across  
 960 conditions:  $r(4) = 0.598, p = 0.210$ ).

961 We suggest two potential interpretations of these findings. First, it is possible that  
 962 some participants are more “malleable” or “adaptable” with respect to how they organize  
 963 incoming information. When presented with list of items sorted along *any* feature dimen-  
 964 sion, they will simply adopt that feature as a dominant dimension for organizing those  
 965 items and subsequent (randomly ordered) items. This flexibility in memory organization  
 966 might afford such participants a memory advantage, explaining their strong recall perfor-



**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.

967 mance. An alternative interpretation is that each participant comes into our study with a  
 968 “preferred” way of organizing incoming information. If they happen to be assigned to an  
 969 order manipulation condition that matches their preferences, then they will appear to be  
 970 “sensitive” to the order manipulation and also exhibit a high degree of carryover in feature  
 971 clustering from early to late lists. These participants might demonstrate strong recall per-  
 972 formance not because of their inherently superior memory abilities, but rather because the  
 973 specific condition they were assigned to happened to be especially easy for them, given  
 974 their pre-experimental tendencies. To help distinguish between these interpretations, we  
 975 designed an *adaptive* experimental condition (see *Adaptive condition*). The primary ma-  
 976 nipulation in the adaptive condition is that participants each experience three key types  
 977 of lists. On *random* lists, words are ordered randomly (as in the feature rich condition).  
 978 On *stabilize* lists, the presentation order is adjusted to be maximally similar to the current  
 979 estimate of the participant’s memory fingerprint (see *Online “fingerprint” analysis*). Third,  
 980 on *destabilize* lists, the presentation order is adjusted to be *minimally* similar to the current



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

estimate of the participant’s memory fingerprint (see *Ordering “stabilize” and “destabilize” lists by an estimated fingerprint*). The orders in which participants experienced each type of list were counterbalanced across participants to help reduce the influence of potential list-order effects. Because the presentation orders on stabilize and destabilize lists are adjusted to best match each participant’s (potentially unique) memory fingerprint, the adaptive condition removes uncertainty about whether participants’ assigned conditions 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 ( $t(59) = -0.249, p = 0.804, d = -0.017, CI =$

996  $[-2.327, 1.578]$ ). Participants also exhibited stronger temporal clustering on stabilize lists,  
 997 relative to random ( $t(59) = 3.428, p = 0.001, d = 0.306, CI = [1.635, 5.460]$ ) and destabi-  
 998 lize ( $t(59) = 4.174, p < 0.001, d = 0.374, CI = [1.964, 6.968]$ ) lists (Fig. 9B). We found no  
 999 reliable differences in temporal clustering for items on random versus destabilize lists  
 1000 ( $t(59) = -0.880, p = 0.382, d = -0.081, CI = [-3.165, 1.127]$ ).

1001 As in the other experimental manipulations, participants in the adaptive condition  
 1002 exhibited substantial variability with respect to their overall memory performance and  
 1003 their clustering tendencies (Fig. 9C). We found that individual participants who exhibited  
 1004 strong temporal clustering scores also tended to recall more items. This held across  
 1005 subjects, aggregating across all list types ( $r(178) = 0.701, p < 0.001, CI = [0.590, 0.789]$ ), and  
 1006 for each list type individually (all  $rs \geq 0.651$ , all  $ps < 0.001$ ). Taken together, the results  
 1007 from the adaptive condition suggest that each participant comes into the experiment with  
 1008 their own unique memory organization tendencies, as characterized by their memory  
 1009 fingerprint. When participants study lists whose items come pre-sorted according to their  
 1010 unique preferences, they tend to remember more and show stronger temporal clustering.

## 1011 Discussion

1012 We asked participants to study and freely recall word lists. The words on each list (and  
 1013 the total set of lists) were held constant across participants. For each word, we considered  
 1014 (and manipulated) two semantic features (category and size) that reflected aspects of the  
 1015 *meanings* of the words, along with two lexicographic features (word length and first letter),  
 1016 which reflected characteristics of the words' *letters*. These semantic and lexicographic  
 1017 features are intrinsic to each word. We also considered and manipulated two additional  
 1018 visual features (color and location) that affected the *appearance* of each studied item, but  
 1019 could be varied independently of the words' identities. Across different experimental



1020 conditions, we manipulated how the visual features varied across words (within each  
1021 list), along with the orders of each list's words. Although the participants' task (verbally  
1022 recalling as many words as possible, in any order, within one minute) remained constant  
1023 across all of these conditions, and although the set of words they studied from each list  
1024 remained constant, our manipulations substantially affected participants' memories. The  
1025 impact of some of the manipulations also affected how participants remembered *future*  
1026 lists that were sorted randomly.

### 1027 **Recap: visual feature manipulations**

1028 We found that participants in our feature rich condition (where we varied words' ap-  
1029 pearances) recalled similar proportions of words to participants in a reduced condition  
1030 (where appearance was held constant across words). However, varying the words' ap-  
1031 pearances led participants to exhibit much more temporal and feature-based clustering.  
1032 This suggests that even seemingly irrelevant elements of our experiences can affect how  
1033 we remember them.

1034 When we held the within-list variability in participants' visual experiences fixed across  
1035 lists (in the feature rich and reduced conditions), they remembered more words from early  
1036 lists than from late lists. For feature rich lists, they also showed stronger clustering for early  
1037 versus late lists. However, when we *varied* participants' visual experiences across lists (in  
1038 the "reduced (early)" and "reduced (late)" conditions), these early versus late accuracy  
1039 and clustering differences disappeared. Abruptly changing how incidental visual features  
1040 varied across words seemed to act as a sort of "event boundary" that partially reset how  
1041 participants processed and remembered post-boundary lists. Within-list clustering also  
1042 increased in these manipulations, suggesting that the "within-event" words were being  
1043 more tightly associated with each other.

1044 When we held the visual features constant during early lists, but then varied words'  
1045 appearances in later lists (i.e., the reduced (early) condition), participants' overall memory  
1046 performance improved. However, this impact was directional: when we *removed* visual  
1047 features from words in late lists that had been present in early lists (i.e., the reduced (late)  
1048 condition), we saw no memory improvement.

### 1049 **Recap: order manipulations**

1050 When we (stochastically) sorted early lists along different feature dimensions, we found  
1051 several impacts on participants' memories. Sorting early lists semantically (by word cat-  
1052 egory) enhanced participants' memories for those lists, but the effects on performance of  
1053 sorting along other feature dimensions were inconclusive. However, each order manipu-  
1054 lation substantially affected how participants *organized* their memories of words from the  
1055 ordered lists. When we sorted lists semantically, participants displayed stronger semantic  
1056 clustering; when we sorted lists lexicographically, they displayed stronger lexicographic  
1057 clustering; and when we sorted lists visually, they displayed stronger visual clustering.  
1058 Clustering along the unmanipulated feature dimensions in each of these cases was un-  
1059 changed.

1060 The order manipulations we examined also appeared to induce, in some cases, a  
1061 tendency to "clump" similar words within a list. This was most apparent on semantically  
1062 ordered lists, where the probability of initiating recall with a given word seemed to follow  
1063 groupings defined by feature change points.

1064 We also examined the impact of early list order manipulations on memory for late  
1065 lists. At the group level, we found little evidence for lingering "carryover" effects of  
1066 these manipulations: participants in the order manipulation conditions showed similar  
1067 memory performance and clustering on late lists to participants in the corresponding

1068 control (feature rich) condition. At the level of individual participants, however, we  
1069 found several meaningful patterns.

1070 Participants who showed stronger feature clustering on early (order-manipulated) lists  
1071 tended to better remember late (randomly ordered) lists. Participants who remembered  
1072 early lists better also tended to show stronger feature clustering (along their condition's  
1073 feature dimension) on late lists (even though the words on those late lists were presented  
1074 in a random order). We also observed some (weaker) carryover effects of temporal cluster-  
1075 ing. Participants who showed stronger feature clustering (along their condition's feature  
1076 dimension) on early lists tended to show stronger temporal clustering on late lists. And  
1077 participants who showed stronger temporal clustering on early lists also tended to show  
1078 stronger feature clustering on late lists. Essentially, these order manipulations appeared to  
1079 affect each participant differently. Some participants were sensitive to our manipulations,  
1080 and those participants' memory performance was impacted more strongly, both for the  
1081 ordered lists and for future (random) lists. Other participants appeared relatively insen-  
1082 sitive to our manipulations, and those participants showed little carryover effects on late  
1083 lists.

1084 These results at the individual participant level suggested to us that either (a) some  
1085 participants were more sensitive to *any* order manipulation, or (b) some participants might  
1086 be more (or less) sensitive to manipulations along *particular* (e.g., preferred) feature dimen-  
1087 sions. To help distinguish between these possibilities, we designed an adaptive condition  
1088 whereby we attempted to manipulate whether participants studied words in an order that  
1089 either matched or mismatched our estimate of how they would cluster or organize the  
1090 studied words in memory (i.e., their idiosyncratic memory fingerprint). We found that  
1091 when we presented words in orders that were consistent with participants' memory fin-  
1092 gerprints, they remembered more words overall and showed stronger temporal clustering.

1093 This comports well with the second possibility described above. Specifically, each partici-  
1094 pant seems to bring into the experiment their own idiosyncratic preferences and strategies  
1095 for organizing the words in their memory. When we presented the words in an order  
1096 consistent with each participant’s idiosyncratic fingerprint, their memory performance  
1097 improved. This might indicate that the participants were spending less cognitive effort  
1098 “reorganizing” the incoming words on those lists, which freed up resources to devote to  
1099 encoding processes instead.

### 1100 **Memory consequences of feature variability**

1101 Several prior studies have examined how varying the richness or experiences, or the  
1102 extensive of encoding, can affect memory. Although specific details differ (Bonin et al.,  
1103 2022), in general these studies have found that richer and more deeply or extensively  
1104 encoded experiences are remembered better (Hargreaves et al., 2012; Madan, 2021; Mein-  
1105 hardt et al., 2020). Our findings help to elucidate an additional factor that may contribute  
1106 to these phenomenon. For example, our finding that participants better remember “fea-  
1107 ture rich” lists (where words’ appearances are varied) than “reduced” lists (where words’  
1108 appearances are held constant) only when those feature rich lists are presented *after* re-  
1109 duced lists suggests that some factors that influence the richness or depth of encoding  
1110 may be relative, rather than absolute. In other words, *increases* in richness (e.g., relative  
1111 to a recency-weighted baseline) may be more important than the overall complexity or  
1112 numbers of features.

1113 Some prior studies have suggested that people can “cue” their memories using different  
1114 “strategies” or “pathways” for searching for the target information. For example, modern  
1115 accounts of free recall typically posit that memory search typically begins by matching  
1116 the current state of mental context with the contexts associated with other items in mem-

1117 ory (Kahana, 2020). Since context is the defining hallmark of episodic memory (Tulving,  
1118 1983), context-based search can be described as an “episodic” pathway to recall. When  
1119 episodic cueing fails to elicit a match, participants may then search for items that are simi-  
1120 lar to the current mental context or mental state along other dimensions, such as semantic  
1121 similarity (Davachi et al., 2003; Socher et al., 2009). These multiple pathways accounts of  
1122 memory search also provide a potential explanation of why participants might have an  
1123 easier time remembering richer stimuli (or experiences): richer stimuli and experiences  
1124 might have more features that could be used to cue memory search. Our work suggests  
1125 that there may be some additional factors at play with respect to the *dynamics* of these pro-  
1126 cesses. In particular, we only observed memory benefits for “richer” stimuli when they  
1127 were encountered after more “impoverished” stimuli (in the reduced (early) condition).  
1128 This suggests that the pathways available to recall a given item may also depend on recent  
1129 prior experiences.

1130 We did *not* find any evidence that changing words’ appearances *harmed* memory per-  
1131 formance, e.g., by distracting them with irrelevant information (Lange, 2005; Marsh et al.,  
1132 2012, 2015; Reinitz et al., 1992). Nor did we find any evidence that *changes* in the presence  
1133 of potentially “distracting” features adversely affected memory. For example, when we  
1134 increased or decreased the variability in words’ appearances on late versus early lists (as in  
1135 the reduced (early) and reduced (late) conditions), we found no evidence that this harmed  
1136 participants’ memories. One potential interpretation under the “multiple pathways to  
1137 recall” framework is that the availability of multiple pathways to recall do not appear to  
1138 specifically interfere with each other.

## 1139 **Context effects on memory performance and organization**

1140 In real-world experience, each moment's unique blend of contextual features (where we  
1141 are, who we are with, what else we are thinking of at the time, what else we experience  
1142 nearby in time, etc.) plays an important role in how we interpret, experience, and re-  
1143 member that moment, and how we relate it to our other experiences (e.g., for review see  
1144 Manning, 2020). What are the analogues of real-world contexts in laboratory tasks like  
1145 the free recall paradigm employed in our study? In general, modern formal accounts of  
1146 free recall (Kahana, 2020) describe context as comprising a mix of (a) features pertaining  
1147 to or associated with each item and (b) other items and thoughts experienced nearby in  
1148 time, e.g., that might still be "lingering" in the participant's thoughts at the time they  
1149 study the item. Item features can include semantic properties (i.e., features related to the  
1150 item's meaning), lexicographic properties (i.e., features related to the item's letters), sen-  
1151 sory properties (i.e., feature related to the item's appearance, sound, smell, etc.), emotional  
1152 properties (i.e., features related to how meaningful the item is, whether the item evokes  
1153 positive or negative feelings, etc.), utility-related properties (e.g., features that describe  
1154 how an item might be used or incorporated into a particular task or situation), and more.  
1155 Essentially any aspect of the participant's experience that can be characterized, measured,  
1156 or otherwise described can be considered to influence the participant's mental context at  
1157 the moment they experience that item. Temporally proximal features include aspects of  
1158 the participant's internal or external experience that are *not* specifically occurring at the  
1159 moment they encounter an item, but that nonetheless influence how they process the item.  
1160 Thoughts related to percepts, goals, expectations, other experiences, and so on that might  
1161 have been cued (directly or indirectly) by the participant's recent experiences prior to the  
1162 current moment all fall into this category. Internally driven mental states, such as thinking  
1163 about an experience unrelated to the experiment, also fall into this category.

1164 Contextual features need not be intentionally or consciously perceived by the partic-  
1165 ipant to affect memory, nor do they need to be relevant to the task instructions or the  
1166 participant's goals. Incidental factors such as font color (Jones and Pyc, 2014), background  
1167 color (Isarida and Isarida, 2007), inter-stimulus images (Chiu et al., 2021; Gershman et al.,  
1168 2013; Manning et al., 2016), background sounds (Sahakyan and Smith, 2014; ?), secondary  
1169 tasks (Masicampo and Sahakyan, 2014; Oberauer and Lewandowsky, 2008; Polyn et al.,  
1170 2009), and more can all impact how participants remember, and organize in memory, lists  
1171 of studied items.

1172 Consistent with this prior work, we found that participants were sensitive to task-  
1173 irrelevant visual features. We also found that changing the dynamics of those task-  
1174 irrelevant visual features (in the reduced (early) and reduced (late) conditions) *also* affected  
1175 participants' memories. This suggests that it is not only the contextual features themselves  
1176 that affect memory, but also the *dynamics* of context—i.e., how the contextual features  
1177 associated with each item change over time.

### 1178 **Priming effects on memory performance and organization**

1179 When our ongoing experiences are ambiguous, we can draw on our past experiences,  
1180 expectations, and other real, perceived, or inferred cues to help resolve these ambiguities.  
1181 We may also be overtly or covertly “primed” to influence how we are likely to resolve  
1182 ambiguities. For example, before listening to a story with several equally plausible inter-  
1183 pretations, providing participants with “background” information beforehand can lead  
1184 them towards one interpretation versus another (Yeshurun et al., 2017). More broadly, our  
1185 conscious and unconscious biases and preferences can influence not only how we interpret  
1186 high-level ambiguities, but even how we process low-level sensory information (Katabi  
1187 et al., 2023).

1188 In more simplified scenarios, like list-learning paradigms, the stimuli and tasks partic-  
1189 ipants encounter before studying a given list can influence what and how they remember.  
1190 For example, when participants are directed to suppress, disregard, or ignore “distracting”  
1191 stimuli early on in an experiment, participants often tend to remember those stimuli less  
1192 well when they are re-used as to-be-remembered targets later on in the experiment (Tip-  
1193 per, 1985). In general, participants’ memories can be influenced by exposing them to  
1194 a wide range of positive and negative priming factors before they encounter the to-be-  
1195 remembered information (Balota et al., 1992; Clayton and Chattin, 1989; Donnelly, 1988;  
1196 Flexser and Tulving, 1982; Gotts et al., 2012; Huang et al., 2004; Huber, 2008; Huber et al.,  
1197 2001; McNamara, 1994; Neely, 1977; Rabinowitz, 1986; Tulving and Schacter, 1991; Watkins  
1198 et al., 1992; Wiggs and Martin, 1998).

1199 The order manipulation conditions in our experiment show that participants can also be  
1200 primed to pick up on more subtle statistical structure in their experiences, like the dynamics  
1201 of how the presentation orders of stimuli vary along particular feature dimensions. These  
1202 order manipulations affected not only how participants remembered the manipulated  
1203 lists, but also how they remembered *future* lists with different (randomized) temporal  
1204 properties.

### 1205 **Free recall of blocked versus random categorized word lists**

1206 A large number of prior studies have compared participants’ memories for categorized  
1207 word lists that are presented in blocked versus random orders. In “blocked” lists, all  
1208 of the words from a given semantic category (e.g., animals) are presented consecutively,  
1209 whereas in “random” lists, the words from different categories are intermixed. Most of  
1210 these studies report that participants tend to better remember blocked (versus random)  
1211 lists (Bower et al., 1969; Cofer et al., 1966; D’Agostino, 1969; Dallett, 1964; Kintsch, 1970;



1212 Luek et al., 1971; Puff, 1974; Shapiro, 1970; ?; ?). Other studies suggest that these order  
1213 effects may also be modulated by factors like list length and the numbers of exemplars in  
1214 each category (e.g., Borges and Mangler, 1972).

1215     Although we did not directly manipulate “blocking” in our order manipulation condi-  
1216 tions, our sorting procedures in those conditions (see *Constructing feature-sorted lists*) have  
1217 *indirect* effects on the lists’ blockiness. For example, lists that are stochastically sorted by  
1218 semantic category will tend to contain runs of several same-category words in succession.  
1219 Consistent with the above work on blocked versus random categorized lists, we found  
1220 that participants tended to better remember lists that were sorted semantically (Fig. 5B).  
1221 However, this memory improvement did not appear to extend to the other order ma-  
1222 nipulation conditions we considered (e.g., to lexicographically or visually sorted lists).  
1223 One possibility is that the memory benefits of blocked versus random lists are specific to  
1224 semantic categories, and do not generalize to other feature dimensions. Another possi-  
1225 bility is that the memory benefits are due to the presence of infrequent “jumps” between  
1226 successive items (e.g., from different categories). Because the features we manipulated in  
1227 the lexicographic and visual conditions were less categorical than the semantic features,  
1228 feature values across words in those conditions tended to vary more gradually. Relatively  
1229 stable features that are punctuated by infrequent large changes (e.g., as words transition  
1230 from a same-category sequence to a new category) may also relate to perceived “event  
1231 boundaries,” which can have important consequences for memory (DuBrow and Davachi,  
1232 2013, 2016; DuBrow et al., 2017; Radvansky and Zacks, 2017).

### 1233 **Expectation, event boundaries, and situation models**

1234 Our findings that participants’ current and future memory behaviors are sensitive to  
1235 manipulations in which features change over time, and how features change across items

1236 and lists, suggest parallels with studies on how we form expectations and predictions,  
1237 segment our continuous experiences into discrete events, and make sense of different  
1238 scenarios and situations. Each of these real-world cognitive phenomena entail identifying  
1239 statistical regularities in our experiences, and exploiting those regularities to gain insight,  
1240 form inferences, organize or interpret memories, and so on. Our past experiences enable  
1241 us to predict what is likely to happen in the future, given what happened “next” in our  
1242 previous experiences that were similar to now (Barron et al., 2020; Brigard, 2012; Chow  
1243 et al., 2016; Eichenbaum and Fortin, 2009; Gluck et al., 2002; Goldstein et al., 2021; Griffiths  
1244 and Steyvers, 2003; Jones and Pashler, 2007; Kim et al., 2014; Manning, 2020; Tamir and  
1245 Thornton, 2018; Xu et al., 2023).

1246 When our expectations are violated, such as when our observations disagree with our  
1247 predictions, we may perceive the “rules” or “situation” to have changed. *Event boundaries*  
1248 denote abrupt changes in the state of our experience, for example, when we transition  
1249 from one situation to another (Radvansky and Zacks, 2017; Zwaan and Radvansky, 1998).  
1250 Crossing an event boundary can impair our memory for pre-boundary information and en-  
1251 hance our memory for post-boundary information (DuBrow and Davachi, 2013; Manning  
1252 et al., 2016; Radvansky and Copeland, 2006; Sahakyan and Kelley, 2002). Event bound-  
1253 aries are also tightly associated with the notion of *situation models* and *schemas*—mental  
1254 frameworks for organizing our understanding about the rules of how we and others are  
1255 likely to behave, how events are likely to unfold over time, how different elements are  
1256 likely to interact, and so on. For example, a situation model pertaining to a particular  
1257 restaurant might set our expectations about what we are likely to experience when we  
1258 visit that restaurant (e.g., what the building will look like, how it will smell when we enter,  
1259 how crowded the restaurant is likely to be, the sounds we are likely to hear, etc.). Similarly,  
1260 as mentioned in the *Introduction*, we might learn a schema describing how events are likely

1261 to unfold *across* any sit-down restaurant—e.g., open the door, wait to be seated, receive a  
1262 menu, decide what to order, place the order, and so on. Situation models and schemas can  
1263 help us to generalize across our experiences, and to generate expectations about how new  
1264 experiences are likely to unfold. When those expectations are violated, we can perceive  
1265 ourselves to have crossed into a new situation.

1266 In our study, we found that abruptly changing the “rules” about how the visual  
1267 appearances of words are determined, or about the orders in which words are presented,  
1268 can lead participants to behave similarly to what one might expect upon crossing an event  
1269 boundary. Adding variability in font color and presentation location for words on late  
1270 lists, after those visual features had been held constant on early lists, led participants to  
1271 remember more words on those later lists. One potential explanation is that participants  
1272 perceive an “event boundary” to have occurred when they encounter the first “late” list.  
1273 According to contextual change accounts of memory across event boundaries (e.g., Flores  
1274 et al., 2017; Gold et al., 2017; Pettijohn et al., 2016; Sahakyan and Kelley, 2002), this could  
1275 help to explain why participants in the reduced (early) condition exhibited better overall  
1276 memory performance. Specifically, their memory for late list items could benefit from less  
1277 interference from early list items, and the contextual features associated with late list items  
1278 (after the “event boundary”) might serve as more specific recall cues for those late items  
1279 (relative to if the boundary had not occurred).

## 1280 **How do different types of clustering relate to each other, and to memory perfor-** 1281 **mance?**

1282 When the words on a studied list are presented in a random order, different types of  
1283 clustering in participants’ recalls often tend to be negatively correlated. For example,  
1284 words that occur nearby on the list will not (on average) tend to be semantically related, and

1285 vice versa. Therefore a participant who shows a strong tendency to temporally cluster their  
1286 recalls will tend to show weaker semantic clustering, and so on (Healey and Uitvlugt, 2019;  
1287 Howard and Kahana, 2002b; Sederberg et al., 2010). Further, there is some evidence that  
1288 temporal clustering is positively correlated with memory performance, whereas semantic  
1289 clustering is negatively correlated with memory performance (Sederberg et al., 2010).

1290 The notion of “multiple pathways to recall” discussed above (see *Memory consequences*  
1291 *of feature variability*) suggests one potential explanation for these patterns. For exam-  
1292 ple, temporal clustering has been proposed to reflect reliance on contextual cues in an  
1293 “episodic” pathway to search memory, whereas semantic clustering reflects a relies on  
1294 specific item features. These two pathways may “compete” with each other during re-  
1295 call (Socher et al., 2009). Meanwhile, extra-list intrusion errors (i.e., false “recalls” of items  
1296 that were never encountered on the list) often tend to share semantic features with recently  
1297 recalled items (Zaromb et al., 2006) and also often lead the participant to stop recalling  
1298 additional items (Miller et al., 2012). Speculatively, over-reliance on semantic cues may  
1299 lead to more intrusion errors, which in turn may lead to fewer recalls overall.

1300 Our findings extend these prior results to consider lists that are *not* ordered randomly.  
1301 Because ordering the words on a list along a particular feature dimension removes the  
1302 “conflict” between temporal and feature clustering, the order manipulation conditions in  
1303 our study represent an “edge case” whereby different pathways to recall are not neces-  
1304 sarily in conflict with each other. For example, the same participants who exhibit strong  
1305 feature clustering *also* show strong temporal clustering on ordered lists (Fig. 7E). This  
1306 is presumably at least partly due to an inability to separate temporal and feature clus-  
1307 tering on ordered lists (also see *Factoring out the effects of temporal clustering*). However,  
1308 features that change gradually with time (i.e., presentation position) could also serve to  
1309 strengthen the episodic (contextual) cues associated with each item. In other words, par-

1310 ticipants might essentially combine multiple noisy measures of change to form a more  
1311 stable internal representation of temporal context.

## 1312 **Theoretical implications**

1313 Although most modern formal theories of episodic memory have been developed and  
1314 tested to explain memory for list-learning tasks (Kahana, 2020), a number of recent studies  
1315 suggest some substantial differences between memory for lists versus naturalistic stim-  
1316 uli (e.g., real-world experiences, narratives, films, etc.; Heusser et al., 2021; Lee et al., 2020;  
1317 Manning, 2021; Nastase et al., 2020). One reason is that naturalistic stimuli are often much  
1318 more engaging than the highly simplified list-learning tasks typically employed in the  
1319 psychological laboratory, perhaps leading participants to pay more attention, exert more  
1320 effort, and stay more consistently motivated to perform well (Nastase et al., 2020). Another  
1321 reason is that the temporal unfoldings of events and occurrences in naturalistic stimuli  
1322 tend to be much more meaningful than the temporal unfoldings of items on typical lists  
1323 used in laboratory memory tasks. Real-world events exhibit important associations at a  
1324 broad range of timescales. For example, an early detail in a detective story may prove to  
1325 be a clue to solving the mystery later on. Further, what happens in one moment typically  
1326 carries some predictive information about what came before or after (Xu et al., 2023). In  
1327 contrast, the lists used in laboratory memory tasks are most often ordered randomly, by  
1328 design, to *remove* meaningful temporal structure in the stimulus (Kahana, 2012).

1329 On one hand, naturalistic stimuli provide a potential means of understanding how our  
1330 memory systems function in the circumstances we most often encounter in our everyday  
1331 lives. This implies that, to understand how memory works in the “real world,” we should  
1332 study memory for stimuli that reflect the relevant statistical structure of real-world expe-  
1333 riences. On the other hand, naturalistic stimuli can be difficult to precisely characterize or

1334 model, making it difficult to distinguish whether specific behavioral trends follow from  
1335 fundamental workings of our memory systems, from some aspect of the stimulus, or from  
1336 idiosyncratic interactions or interference between participants' memory systems and the  
1337 stimulus. This challenge implies that, to understand the fundamental nature of memory  
1338 in its "pure" form, we should study memory for highly simplified stimuli that can pro-  
1339 vide relatively unbiased (compared with real-world experiences) measures of the relevant  
1340 patterns and tendencies.

1341     The experiment we report in this paper was designed to help bridge some of this gap  
1342 between naturalistic tasks and more traditional list-learning tasks. We had people study  
1343 word lists similar to those used in classic memory studies, but we also systematically var-  
1344 ied the lists' "richness" (by adding or removing visual features) and temporal structure  
1345 (through order manipulations that varied over time and across experimental conditions).  
1346 We found that participants' memory behaviors were sensitive to these manipulations.  
1347 Some of the manipulations led to changes that were common across people (e.g., more  
1348 temporal clustering when words' appearances were varied, enhanced memory for lists  
1349 following an "event boundary," more feature clustering on order-manipulated lists, etc.).  
1350 Other manipulations led to changes that were idiosyncratic (especially carryover effects  
1351 from order manipulations; e.g., participants who remembered more words on early order-  
1352 manipulated lists tended to show stronger feature clustering for their condition's feature  
1353 dimension on late randomly ordered lists, etc.). We also found that participants remem-  
1354 bered more words from lists that were sorted to align with their idiosyncratic clustering  
1355 preferences. Taken together, our results suggest that our memories are susceptible to ex-  
1356 ternal influences (i.e., to the statistical structure of ongoing experiences), but the effects of  
1357 past experiences on future memory are largely idiosyncratic across people.

## 1358 **Potential applications**

1359 Every participant in our study encountered exactly the same words, split into exactly the  
1360 same lists. But participants' memory performance, the orders in which they recalled the  
1361 words, and the effects of early list manipulations on later lists all varied according to how  
1362 we presented the to-be-remembered words.

1363 Our findings raise a number of exciting questions. For example, how far might these  
1364 manipulations be extended? In other words, might there be more sophisticated or clever  
1365 feature or order manipulations that one could implement to have stronger impacts on  
1366 memory? Are there limits to how much impact (on memory performance and/or or-  
1367 ganization) these sorts of manipulations can have? Are those limits universal across  
1368 people, or are there individual differences (based on prior experiences, natural strate-  
1369 gies, neuroanatomy, etc.) that impose person-specific limits on the potential impact of  
1370 presentation-level manipulations on memory?

1371 Our findings indicate that the ways word lists are presented affects how people re-  
1372 member them. To the extent that word list memory reflects memory processes that are  
1373 relevant to real-world experiences, one could imagine potential real-world applications of  
1374 our findings. For example, we found that participants remembered more words when the  
1375 presentation order agreed with their memory fingerprints. If analogous fingerprints could  
1376 be estimated for classroom content, perhaps they could be utilized manually by teachers,  
1377 or even by automated content-presentation systems, to optimize how and what students  
1378 remember.

## 1379 **Concluding remarks**

1380 Our work raises deep questions about the fundamental nature of human learning. What  
1381 are the limits of our memory systems? How much does what we remember (and how we

remember) depend on how we learn or experience the to-be-remembered content? We know that our expectations, strategies, situation models learned through prior experiences, and more collectively shape how our experiences are remembered. But those aspects of our memory are not fixed: when we are exposed to the same experience in a new way, it can change how we remember that experience, and also how we remember, process, or perceive *future* experiences.

### Author contributions

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

### Author note

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



## 1405 Acknowledgements

1406 We acknowledge useful discussions, assistance in setting up an earlier (unpublished)  
1407 version of this study, and assistance with some of the data collection efforts from Rachel  
1408 Chacko, Joseph Finkelstein, Sheherzad Mohyidin, Lucy Owen, Gal Perlman, Jake Rost,  
1409 Jessica Tin, Marisol Tracy, Peter Tran, and Kirsten Ziman. Our work was supported in part  
1410 by NSF CAREER Award Number 2145172 to JRM. The content is solely the responsibility  
1411 of the authors and does not necessarily represent the official views of our supporting  
1412 organizations. The funders had no role in study design, data collection and analysis,  
1413 decision to publish, or preparation of the manuscript.

## 1414 References

- 1415 Anderson, J. R. and Bower, G. H. (1972). Recognition and retrieval processes in free recall.  
1416 *Psychological Review*, 79(2):97–123.
- 1417 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its  
1418 control processes. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning*  
1419 *and Motivation*, volume 2, pages 89–105. Academic Press, New York, NY.
- 1420 Baddeley, A. D. (1968). Prior recall of newly learned items and the recency effect in free  
1421 recall. *Canadian Journal of Psychology*, 22:157–163.
- 1422 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event  
1423 schemas during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.
- 1424 Balota, D. A., Black, S. R., and Cheney, M. (1992). Automatic and attentional priming in  
1425 young and older adults: reevaluation of the two-process model. *Journal of Experimental*  
1426 *Psychology: Human Perception and Performance*, 18(2):485–502.

- 1427 Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a  
1428 predictive coding account. *Progress in Neurobiology*, 192:101821–101834.
- 1429 Bonin, P., Thiebaut, G., Bugaiska, A., and Méot, A. (2022). Mixed evidence for a richness-of-  
1430 encoding account of animacy effects in memory from the generation-of-ideas paradigm.  
1431 *Current Psychology*, 41:1653–1662.
- 1432 Borges, M. A. and Mangler, G. (1972). Effect of within-category spacing on free recall.  
1433 *Journal of Experimental Psychology*, 92(2):207–214.
- 1434 Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged  
1435 associates. *Journal of General Psychology*, 49:229–240.
- 1436 Bousfield, W. A., Sedgewick, C. H., and Cohen, B. H. (1954). Certain temporal character-  
1437 istics of the recall of verbal associates. *American Journal of Psychology*, 67:111–118.
- 1438 Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive*  
1439 *Psychology*, 11(2):177–220.
- 1440 Bower, G. H., Lesgold, A. M., and Tieman, D. (1969). Grouping operations in free recall.  
1441 *Journal of Verbal Learning and Verbal Behavior*, 8(4):481–493.
- 1442 Brigard, F. D. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*,  
1443 3(420):1–3.
- 1444 Chiu, Y.-C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., and Sahakyan, L. (2021). Sepa-  
1445 ration of item and context in item-method directed forgetting. *NeuroImage*, 235:117983.
- 1446 Chow, W.-Y., Momma, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory  
1447 retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.

- 1448 Clayton, K. and Chattin, D. (1989). Spatial and semantic priming effects in tests of spa-  
1449 tial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,  
1450 15(3):495–506.
- 1451 Clewett, D., DuBrow, S., and Davachi, L. (2019). Transcending time in the brain: how  
1452 event memories are constructed from experience. *Hippocampus*, 29(3):162–183.
- 1453 Cofer, C. N., Bruce, D. R., and Reicher, G. M. (1966). Clustering in free recall as a function  
1454 of certain methodological variations. *Journal of Experimental Psychology: General*, 71:858–  
1455 866.
- 1456 D’Agostino, P. R. (1969). The blocked-random effect in recall and recognition. *Journal of*  
1457 *Verbal Learning and Verbal Behavior*, 8:815–820.
- 1458 Dallett, K. M. (1964). Number of categories and category information in free recall. *Journal*  
1459 *of Experimental Psychology*, 68:1–12.
- 1460 Darley, C. F. and Murdock, B. B. (1971). Effects of prior free recall testing on final recall  
1461 and recognition. *Journal of Experimental Psychology: General*, 91:66–73.
- 1462 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct  
1463 medial temporal lobe processes build item and source memories. *Proceedings of the*  
1464 *National Academy of Sciences, USA*, 100(4):2157–2162.
- 1465 Donnelly, R. E. (1988). Priming effects in successive episodic tests. *Journal of Experimental*  
1466 *Psychology: Learning, Memory, and Cognition*, 14:256–265.
- 1467 Drewnowski, A. and Murdock, B. B. (1980). The role of auditory features in memory span  
1468 for words. *Journal of Experimental Psychology: Human Learning and Memory*, 6:319–332.

- 1469 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for  
1470 the sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–  
1471 1286.
- 1472 DuBrow, S. and Davachi, L. (2016). Temporal binding within and across events. *Neurobi-*  
1473 *ology of Learning and Memory*, 134:107–114.
- 1474 DuBrow, S., Rouhani, N., Niv, Y., and Norman, K. A. (2017). Does mental context drift or  
1475 shift? *Current Opinion in Behavioral Sciences*, 17:141–146.
- 1476 Eichenbaum, H. and Fortin, N. J. (2009). The neurobiology of memory based predictions.  
1477 *Philosophical Transactions of the Royal Society of London Series B*, 364(1521):1183–1191.
- 1478 Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological*  
1479 *Review*, 62:145–154.
- 1480 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory?  
1481 *Psychological Science*, 22(2):243–252.
- 1482 Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall:  
1483 a challenge for unitary models of recency. *Journal of Experimental Psychology: Learning,*  
1484 *Memory, and Cognition*, 36:324–347.
- 1485 Farrell, S. (2014). Correcting the correction of conditional recency slopes. *Psychonomic*  
1486 *Bulletin and Review*, 21:1174–1179.
- 1487 Flexser, A. J. and Tulving, E. (1982). Priming and recognition failure. *Journal of Verbal*  
1488 *Learning and Verbal Behavior*, 21:237–248.
- 1489 Flores, S., Bailey, H. R., Eisenberg, M. L., and Zacks, J. M. (2017). Event segmentation

improves event memory up to one month later. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(8):1183.

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

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

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

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

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

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

- 1512 Griffiths, T. L. and Steyvers, M. (2003). Prediction and semantic association. *Advances in*  
1513 *Neural Information Processing Systems*, 15.
- 1514 Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml,  
1515 M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages  
1516 2338–2342.
- 1517 Hargreaves, I. S., Pexman, P. M., Johnson, J. C., and Zdravilova, L. (2012). Richer concepts  
1518 are better remembered: number of features effects in free recall. *Frontiers in Human*  
1519 *Neuroscience*, 6:doi.org/10.3389/fnhum.2012.00073.
- 1520 Healey, M. K. and Uitvlugt, M. G. (2019). The role of control processes in temporal and  
1521 semantic contiguity. *Memory and Cognition*, 47:719–737.
- 1522 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail:  
1523 a Python toolbox for analyzing and plotting free recall data. *Journal of Open Source*  
1524 *Software*, 10.21105/joss.00424.
- 1525 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal  
1526 behavioral and neural signatures of transforming experiences into memories. *Nature*  
1527 *Human Behavior*, 5:905–919.
- 1528 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). HyperTools: a  
1529 Python toolbox for gaining geometric insights into high-dimensional data. *Journal of*  
1530 *Machine Learning Research*, 18(152):1–6.
- 1531 Hogan, R. M. (1975). Interitem encoding and directed search in free recall. *Memory and*  
1532 *Cognition*, 3:197–209.
- 1533 Howard, M. W. and Kahana, M. J. (1999). Contextual variability and serial position effects

- 1534 in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:923–  
1535 941.
- 1536 Howard, M. W. and Kahana, M. J. (2002a). A distributed representation of temporal  
1537 context. *Journal of Mathematical Psychology*, 46:269–299.
- 1538 Howard, M. W. and Kahana, M. J. (2002b). When does semantic similarity help episodic  
1539 retrieval? *Journal of Memory and Language*, 46:85–98.
- 1540 Huang, L., Holcombe, A. O., and Pashler, H. (2004). Repetition priming in visual search:  
1541 episodic retrieval, not feature priming. *Memory and Cognition*, 32:12–20.
- 1542 Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental*  
1543 *Psychology: General*, 137(2):324–347.
- 1544 Huber, D. E., Shiffrin, R. M., Lyle, K. B., and Ruys, K. I. (2001). Perception and preference  
1545 in short-term word priming. *Psychological Review*, 108(1):149–182.
- 1546 Isarida, T. and Isarida, T. K. (2007). Environmental context effects of background color in  
1547 free recall. *Memory and Cognition*, 35(7):1620–1629.
- 1548 Jenkins, J. J. and Russell, W. A. (1952). Associative clustering during recall. *Journal of*  
1549 *Abnormal and Social Psychology*, 47:818–821.
- 1550 Jones, A. C. and Pyc, M. A. (2014). The production effect: costs and benefits in free recall.  
1551 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):300–305.
- 1552 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing  
1553 prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 1554 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*,  
1555 24:103–109.

- 1556 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York,  
1557 NY.
- 1558 Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychol-*  
1559 *ogy*, 71:107–138.
- 1560 Kahana, M. J., Howard, M. W., and Polyn, S. M. (2008). Associative processes in episodic  
1561 memory. In Roediger III, H. L., editor, *Cognitive Psychology of Memory*, pages 476–490.  
1562 Elsevier, Oxford, UK.
- 1563 Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., and Yeshurun, Y. (2023). Deeper than  
1564 you think: partisanship-dependent brain responses in early sensory and motor brain  
1565 regions. *The Journal of Neuroscience*, pages doi.org/10.1523/JNEUROSCI.0895–22.2022.
- 1566 Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning  
1567 of memories by context-based prediction error. *Proceedings of the National Academy of*  
1568 *Sciences, USA*, In press.
- 1569 Kimball, D. R., Smith, T. A., and Kahana, M. J. (2007). The fSAM model of false recall.  
1570 *Psychological Review*, 114(4):954–993.
- 1571 Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 1572 Lange, E. B. (2005). Disruption of attention by irrelevant stimuli in serial recall. *Journal of*  
1573 *Memory and Language*, 43(4):513–531.
- 1574 Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases  
1575 of human memory. *Current Opinion in Behavioral Sciences*, 32:111–119.
- 1576 Lohnas, L. J., Polyn, S. M., and Kahana, M. J. (2010). Modeling intralist and interlist effects  
1577 in free recall. In *Psychonomic Society*, Saint Louis, MO.



- 1578 Luek, S. P., McLaughlin, J. P., and Cicala, G. A. (1971). Effects of blocking of input and  
1579 blocking of retrieval cues on free recall learning. *Journal of Experimental Psychology*,  
1580 91(1):159–161.
- 1581 Madan, C. R. (2021). Exploring word memorability: how well do different word properties  
1582 explain item free-recall probability? *Psychonomic Bulletin and Review*, 28:583–595.
- 1583 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
1584 *Handbook of Human Memory*. Oxford University Press.
- 1585 Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”  
1586 function? *Psychological Review*, 128(4):711–725.
- 1587 Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. A.  
1588 (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic*  
1589 *Bulletin and Review*, 23(5):1534–1542.
- 1590 Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free  
1591 recall. *Memory*, 20(5):511–517.
- 1592 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic  
1593 memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.
- 1594 Manning, J. R., Notaro, G. M., Chen, E., and Fitzpatrick, P. C. (2022). Fitness tracking  
1595 reveals task-specific associations between memory, mental health, and physical activity.  
1596 *Scientific Reports*, 12(13822):doi.org/10.1038/s41598-022-17781-0.
- 1597 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory pat-  
1598 terns in temporal lobe reveal context reinstatement during memory search. *Proceedings*  
1599 *of the National Academy of Sciences, USA*, 108(31):12893–12897.

- 1600 Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., and Kahana, M. J. (2012).  
 1601 Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clus-  
 1602 tering during memory search. *The Journal of Neuroscience*, 32(26):8871–8878.
- 1603 Marsh, J. E., Beaman, C. P., Hughes, R. W., and Jones, D. M. (2012). Inhibitory control in  
 1604 memory: evidence for negative priming in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5):1377–1388.
- 1606 Marsh, J. E., Sörqvist, P., Hodgetts, H. M., Beaman, C. P., and Jones, D. M. (2015). Distraction  
 1607 control processes in free recall: benefits and costs to performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(1):118–133.
- 1609 Masicampo, E. J. and Sahakyan, L. (2014). Imagining another context during encoding off-  
 1610 sets context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6):1772–1777.
- 1612 Masís-Obando, R., Norman, K. A., and Baldassano, C. (2022). Scheme representations in  
 1613 distinct brain networks support narrative memory during encoding and retrieval. *eLife*, 11:e70445.
- 1615 McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:507–520.
- 1617 Meinhardt, M. J., Bell, R., Buchner, A., and Röer, J. P. (2020). Adaptive memory: is  
 1618 the animacy effect on memory due to richness of encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(3):416–426.
- 1620 Miller, J. F., Kahana, M. J., and Weidemann, C. T. (2012). Recall termination in free recall.  
 1621 *Memory and Cognition*, 40(4):540–550.

- 1622 Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman,  
1623 S. J. (2017). The successor representation in human reinforcement learning. *Nature*  
1624 *Human Behavior*, 1:680–692.
- 1625 Moran, R. and Goshen-Gottstein, Y. (2014). The conditional-recency dissociation is con-  
1626 founded with nominal recency: should unitary models of memory still be devaluated?  
1627 *Psychonomic Bulletin and Review*, 21:332–343.
- 1628 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental*  
1629 *Psychology: General*, 64:482–488.
- 1630 Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy  
1631 of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1632 Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhi-  
1633 bitionless spreading activation and limited-capacity attention. *Journal of Experimental*  
1634 *Psychology: General*, 106(3):226–254.
- 1635 Oberauer, K. and Lewandowsky, S. (2008). Forgetting in immediate serial recall: decay,  
1636 temporal distinctiveness, or interference? *Psychological Review*, 115(3):544–576.
- 1637 Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., and Radvansky, G. A.  
1638 (2016). Event boundaries and memory improvement. *Cognition*, 148:136–144.
- 1639 Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of  
1640 context. *Trends in Cognitive Sciences*, 12:24–30.
- 1641 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). Task context and organization in  
1642 free recall. *Neuropsychologia*, 47:2158–2163.

- 1643 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly*  
1644 *Journal of Experimental Psychology*, 17:132–138.
- 1645 Puff, C. R. (1974). A consolidated theoretical view of stimulus-list organization effects in  
1646 free recall. *Psychological Reports*, 34:275–288.
- 1647 Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of  
1648 associative memory. In Bower, G. H., editor, *The Psychology of Learning and Motivation: Advances in Research and Theory*, volume 14, pages 207–262. Academic Press, New York,  
1649 NY.  
1650
- 1651 Rabinowitz, J. C. (1986). Priming in episodic memory. *Journal of Gerontology*, 41:204–213.
- 1652 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:  
1653 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.
- 1654 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition.  
1655 *Current Opinion in Behavioral Sciences*, 17:133–140.
- 1656 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior.  
1657 *Nature Reviews Neuroscience*, 13:713–726.
- 1658 Reinitz, M. T., Lammers, W. J., and Cochran, B. P. (1992). Memory-conjunction errors:  
1659 miscombination of stored stimulus features can produce illusions of memory. *Memory*  
1660 *and Cognition*, 20:1–11.
- 1661 Rissman, J., Eliassen, J. C., and Blumstein, S. E. (2003). An event-related fMRI investigation  
1662 of implicit semantic priming. *Journal of Cognitive Neuroscience*, 15(8):1160–1175.
- 1663 Romney, A. K., Brewer, D. D., and Batchelder, W. H. (1993). Predicting clustering from  
1664 semantic structure. *Psychological Science*, 4:28–34.

- 1665 Sahakyan, L. and Kelley, C. M. (2002). A contextual change account of the directed  
1666 forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,  
1667 28(6):1064–1072.
- 1668 Sahakyan, L. and Smith, J. R. (2014). A long time ago, in a context far, far away: Retro-  
1669 spective time estimates and internal context change. *Journal of Experimental Psychology:*  
1670 *Learning, Memory, and Cognition*, 40(1):86–93.
- 1671 Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclo-*  
1672 *pedic Reference*, 3:501–506.
- 1673 Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008). A context-based theory of  
1674 recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.
- 1675 Sederberg, P. B., Miller, J. F., Howard, W. H., and Kahana, M. J. (2010). The tempo-  
1676 ral contiguity effect predicts episodic memory performance. *Memory and Cognition*,  
1677 38(6):689–699.
- 1678 Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of  
1679 time. *Neural Computation*, 24:134–193.
- 1680 Shapiro, S. I. (1970). Isolation effects, free recall, and organization. *Journal of Psychology*,  
1681 24:178–183.
- 1682 Sirotin, Y. B., Kimball, D. R., and Kahana, M. J. (2005). Going beyond a single list: modeling  
1683 the effects of prior experience on episodic free recall. *Psychonomic Bulletin and Review*,  
1684 12(5):787–805.
- 1685 Slamecka, N. J. and Barlow, W. (1979). The role of semantic and surface features in word  
1686 repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 18:617–627.

- 1687 Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: a review and  
1688 meta-analysis. *Psychonomic Bulletin and Review*, 8(2):203–220.
- 1689 Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009). A  
1690 Bayesian analysis of dynamics in free recall. *Advances in Neural Information Processing*  
1691 *Systems*, 22.
- 1692 Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., and Zacks, J. M. (2011).  
1693 Changes in events alter how people remember recent information. *Journal of Cognitive*  
1694 *Neuroscience*, 23(5):1052–1064.
- 1695 Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception  
1696 affect memory encoding and updating. *Journal of Experimental Psychology: General*,  
1697 138(2):236–257.
- 1698 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in*  
1699 *Cognitive Sciences*, 22(3):201–212.
- 1700 Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *The*  
1701 *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37:571–  
1702 590.
- 1703 Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P.,  
1704 and Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, 316(5821):76–  
1705 82.
- 1706 Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press, New York, NY.
- 1707 Tulving, E. and Schacter, D. L. (1991). Priming and human memory systems. *Science*,  
1708 247:301–305.

- 1709 Watkins, P. C., Mathews, A., Williamson, D. A., and Fuller, R. D. (1992). Mood-congruent  
1710 memory in depression: emotional priming or elaboration? *Journal of Abnormal Psychol-*  
1711 *ogy*, 101(3):581–586.
- 1712 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American*  
1713 *Journal of Psychology*, 35:396–401.
- 1714 Whitely, P. L. (1927). The dependence of learning and recall upon prior intellectual activi-  
1715 ties. *Journal of Experimental Psychology: General*, 10:489–508.
- 1716 Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming.  
1717 *Current Opinion in Neurobiology*, 8(2):227–233.
- 1718 Xu, X., Zhu, Z., and Manning, J. R. (2023). The psychological arrow of time drives  
1719 temporal asymmetries in retrodicting versus predicting narrative events. *PsyArXiv*,  
1720 page doi.org/10.31234/osf.io/yp2qu.
- 1721 Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., and Hasson, U.  
1722 (2017). Same story, different story: the neural representation of interpretive frameworks.  
1723 *Psychological Science*, 28(3):307–319.
- 1724 Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., and  
1725 Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal*  
1726 *of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792–804.
- 1727 Zhang, Q., Griffiths, T. L., and Norman, K. A. (2023). Optimal policies for free recall.  
1728 *Psychological Review*, 130(4):1104–1125.
- 1729 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018).  
1730 Is automatic speech-to-text transcription ready for use in psychological experiments?  
1731 *Behavior Research Methods*, 50:2597–2605.

- 1732 Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation  
1733 models in narrative comprehension: an event-indexing model. *Psychological Science*,  
1734 6(5):292–297.
- 1735 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension  
1736 and memory. *Psychological Bulletin*, 123(2):162–185.