

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 [Labs](#)

\*Corresponding author: [jeremy.r.manning@dartmouth.edu](mailto:jeremy.r.manning@dartmouth.edu)

4 **Abstract**

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

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

## 19 Introduction

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

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

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

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

50 Like remembering real-world experiences, remembering words on a studied list re-  
51 quires distinguishing the current list from the rest of one's experience. To model this fun-  
52 damental memory capability, cognitive scientists have posited a special context representa-  
53 tion that is associated with each list. According to early theories ~~(e.g. Anderson and Bower, 1972; Estes, 1955)-~~  
54 (e.g. Anderson and Bower, 1972; Estes, 1955) context representations are composed of many  
55 features which fluctuate from moment to moment, slowly drifting through a multidimen-  
56 sional feature space. During recall, this representation forms part of the retrieval cue,  
57 enabling us to distinguish list items from non-list items. Understanding the role of context  
58 in memory processes is particularly important in self-cued memory tasks, such as free  
59 recall, where the retrieval cue is "context" itself (Howard and Kahana, 2002a). Conceptu-  
60 ally, the same general processes might be said to describe how real-world contexts evolve  
61 during natural experiences. However, this is still an open area of study (Manning, 2020,  
62 2021).

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

69 and, to a lesser extent, from the beginning of the ~~study~~studied list), as well as semantic  
70 and temporal clustering effects (Howard and Kahana, 2002b; Kahana et al., 2008). The  
71 contiguity effect is an example of temporal clustering, which is perhaps the dominant form  
72 of organization in free recall. This effect can be seen in people's tendencies to successively  
73 recall items that occupied neighboring positions in the studied list (Kahana, 1996). There  
74 are also striking effects of semantic clustering (Bousfield, 1953; Bousfield et al., 1954; Jenk-  
75 ins and Russell, 1952; Manning and Kahana, 2012; Romney et al., 1993), whereby the recall  
76 of a given item is more likely to be followed by recall of a similar or related item than a  
77 dissimilar or unrelated one. In general, people organize memories for words along a wide  
78 variety of stimulus dimensions. ~~As formalized by~~ According to models like the *Context*  
79 *Maintenance and Retrieval* ~~Model~~ model (Polyn et al., 2009), the stimulus features associated  
80 with each word (e.g., the word's meaning, size of the object the word represents, ~~the~~ letters  
81 that make up the word, font size, font color, location on the screen, etc.) are incorporated  
82 into the participant's mental context representation (Manning, 2020; Manning et al., 2015,  
83 2011, 2012; Smith and Vela, 2001). During a memory test, any of these features may serve  
84 as a memory cue, which in turn leads the participant to ~~recall in succession~~ successively  
85 recall words that share stimulus features.

86 A key mystery is whether (and how) the sorts of situation models and schemas that  
87 people use to organize their memories of real-world experiences might map onto the  
88 clustering effects that reflect how people organize their memories for word lists. On  
89 one hand, both situation models and clustering effects reflect statistical regularities in  
90 ongoing experiences. Our memory systems exploit these regularities when generating  
91 inferences about the unobserved past and yet-to-be-experienced future (Bower et al., 1979;  
92 Momennejad et al., 2017; Ranganath and Ritchey, 2012; Schapiro and Turk-Browne, 2015;  
93 Xu et al., 2023). On the other hand, the rich structures of real-world experiences and other

94 naturalistic stimuli that enable people to form deep and meaningful situation models and  
95 schemas have no obvious analogs in simple word lists. Often, lists in free recall studies are  
96 explicitly *designed* to be devoid of exploitable temporal structure, for example ~~by~~ sorting  
97 the words in a random order (Kahana, 2012).

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

119 were presented agreed or disagreed with how each participant preferred to organize their  
120 memories of the studied items ~~versus the orders in which the items were presented.~~

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

## 133 **Materials and methods**

### 134 **Participants**

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

141 Participants ~~either received~~ received either course credit or a one-time \$10 cash pay-

142 ment for enrolling in our study. We asked each participant to fill out a demographic survey  
143 that included questions about their age, gender, ethnicity, race, education, vision, reading  
144 impairments, medications ~~or~~ and recent injuries, coffee consumption on the day of testing,  
145 and level of alertness at the time of testing. All components of the demographics survey  
146 were optional. One participant elected not to fill out any part of the demographic survey,  
147 and all other participants answered some or all of the survey questions.

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

159 Participants were assigned to experimental conditions based loosely on their date of  
160 participation. (This aspect of our procedure helped us to more easily synchronize the  
161 experiment databases across multiple testing computers.) Of the 490 participants who  
162 opted to fill out the demographics survey, reported ages ranged from 17 to 31 years (mean:  
163 19.1 years; standard deviation: 1.356 years). A total of 318 participants reported their  
164 gender as female, 170 reported their gender as male, and two participants declined to  
165 report their gender. A total of 442 participants reported their ethnicity as “not Hispanic or  
166 Latino,” 39 reported their ethnicity as “Hispanic or Latino,” and nine declined to report

167 their ethnicity. Participants reported their races as White (345 participants), Asian (120  
168 participants), Black or African American (31 participants), American Indian or Alaska  
169 Native (11 participants), Native Hawaiian or Other Pacific Islander (four participants),  
170 Mixed race (three participants), Middle Eastern (one participant), and Arab (one partic-  
171 ipant). A total of five participants declined to report their race. We note that several  
172 participants reported more than one of the above racial categories. Participants reported  
173 their highest degrees achieved as “Some college” (359 participants), “High school gradu-  
174 ate” (117 participants), “College graduate” (seven participants), “Some high school” (five  
175 participants), “Doctorate” (one participant), and “Master’s degree” (one participant). A  
176 total of 482 participants reported no reading impairments, ~~and~~ eight reported having  
177 mild reading impairments. A total of 489 participants reported having normal color vi-  
178 sion and one participant reported ~~that they were red-green-color-blind~~ having impaired  
179 color vision. A total of 482 participants reported taking no prescription medications and  
180 having no recent injuries; four participants reported having ADHD, one reported having  
181 dyslexia, one reported having allergies, one reported a recently torn ACL/MCL, and one  
182 reported a concussion from several months prior. The participants reported ~~consuming~~  
183 having consumed 0–3 cups of coffee ~~prior to the~~ on the day of the testing session (mean:  
184 0.32 cups; standard deviation: 0.58 cups). Participants reported their current level of alert-  
185 ness, and we converted their responses to numerical scores as follows: “very sluggish”  
186 (-2), “a little sluggish” (-1), “neutral” (0), “a little alert” (1), and “very alert” (2). Across  
187 all participants, the full range of alertness levels were reported (range: -2–2; mean: 0.35;  
188 standard deviation: 0.89).

189 We dropped from our dataset the one participant who reported having abnormal color  
190 vision, as well as 38 participants whose data were corrupted due to technical failures while  
191 running the experiment or during the daily database merges. In total, this left usable data



192 from 452 participants, broken down by experimental condition as follows: ~~feature-rich~~  
193 ~~feature-rich~~ (67 participants), reduced (61 participants), reduced (early) (42 participants),  
194 reduced (late) (41 participants), category (30 participants), size (30 participants), length (30  
195 participants), first letter (30 participants), color (31 participants), location (30 participants),  
196 and adaptive (60 participants). The participant who declined to fill out their demographic  
197 survey participated in the location condition, and we verified verbally that they had  
198 normal color vision and no significant reading impairments.

## 199 **Experimental design**

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



**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-one-minute verbal recall interval. The labels on the right (and corresponding groupings on the left) denote experimental condition labels.

## 214 Stimuli

215 The stimuli in our paradigm were 256 English words selected in a previous study (Ziman  
216 et al., 2018). ~~The words all~~ All words referred to concrete nouns, and were chosen from  
217 15 unique semantic categories: body parts, building-related, cities, clothing, countries,  
218 flowers, fruits, insects, instruments, kitchen-related, mammals, (US) states, tools, trees,  
219 and vegetables. We also tagged each word according to the approximate size of the object  
220 ~~the word it~~ referred to. Words were labeled as “small” if the corresponding referent object  
221 was likely able to “fit in a standard shoebox” or “large” if the object was larger than  
222 a shoebox. Most semantic categories comprised words that reflected both “small” and  
223 “large” object sizes, but several included only one or the other (e.g., all countries, US states,  
224 and cities are larger than a shoebox; mean number of different sizes per category: 1.33;  
225 standard deviation: 0.49). The ~~numbers number~~ of words in each semantic category also  
226 varied from 12–28 (mean number of words per category: 17.07; standard deviation ~~number~~  
227 ~~of words~~: 4.65). We also identified lexicographic features for each word, including ~~the~~  
228 ~~words’ first letters and lengths~~ its first letter and length (i.e., number of letters). Across  
229 all categories, all possible first letters were represented except for ‘Q’ (average number of  
230 unique first letters per category: ~~11~~ 11.00; standard deviation: ~~2~~ 2.00 letters). Word lengths  
231 ranged from 3–12 letters (average: 6.17 letters; standard deviation: 2.06 letters).

232 We assigned the categorized words into a total of 16 lists with several constraints. First,  
233 we required that each list ~~contained words from~~ contain exactly four unique ~~categories,~~  
234 ~~each with exactly four exemplars from each category~~ words from each of four unique  
235 categories. Second, we required that ~~(across all words on the list)~~ each list contain at  
236 least one ~~instance of both object sizes were represented~~ word representing each of the two  
237 object sizes (“small” and “large”). On average, each category was represented in 4.27 lists  
238 (standard deviation: 1.16 lists). Aside from these two constraints, we randomly assigned

239 each word to a ~~unique list~~ single list (i.e., such that no words appeared in multiple lists  
240 or were omitted entirely). After random assignment, each list contained words with an  
241 average of 11.13 unique starting letters (standard deviation: 1.15 letters) and an average  
242 ~~word~~ length of 6.17 letters (standard deviation: 0.34 letters).

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

253 To select a random font color for each word, we drew three integers uniformly and at  
254 random from the interval ~~{0,255}~~ [0, 254], corresponding to the red (r), green (g), and blue  
255 (b) color channels for that word. To assign random presentation locations to each word,  
256 we selected two floating point numbers uniformly and at random (one for the word's  
257 horizontal  $x$ -coordinate and the other for its vertical  $y$ -coordinate). The bounds of these  
258 coordinates were selected to cover the entire visible area of the display without cutting off  
259 any part of the words. The words were shown on 27-in (diagonal) Retina 5K iMac displays  
260 (resolution: 5120 ~~x~~ by 2880 pixels).

261 Most of the experimental manipulations we carried out entailed presenting or sorting  
262 the presented words differently on the first eight lists participants studied (which we call  
263 "early" lists) versus on the final eight lists they studied ("late" lists). Since every participant

264 studied exactly 16 lists, every list was either “early” or “late” depending on its order in  
265 the list study sequence. (In other words, the “early” and “late” labels capture all of the  
266 lists participants studied.)

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

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

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

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

286 We also designed two conditions ~~where~~in which we varied the words’ visual appear-

ances across lists. In the *reduced (early)* condition, we followed the “reduced” procedure (presenting each word in black at the center of the screen) for early lists, and followed the “~~feature-rich~~feature-rich” procedure (presenting each word in a random color and location) for late lists. Finally, in the *reduced (late)* condition, we followed the ~~feature-rich~~feature-rich procedure for early lists and the reduced procedure for late lists.

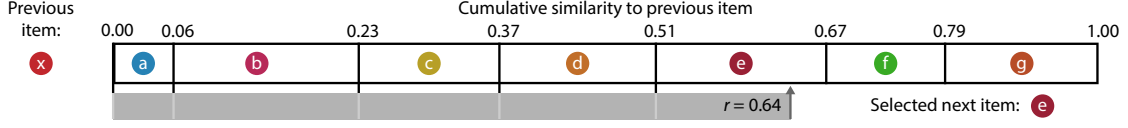
## Order manipulation conditions (Fig. 1, middle six rows)

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

**Defining feature-based distances.** Sorting words according to a given relevant feature requires first defining a distance function for quantifying the dissimilarity between the values of that feature for each pair of ~~features~~words. This function varied according to the type of feature under consideration. Semantic features (category and size) are *categorical*. For these features, we defined a binary distance function: two words were considered to “match” (i.e., have a distance of 0) if their labels were the same (i.e., both from the same semantic category or both of the same size). If two words’ labels were different for a given feature, we defined the words to have a distance of ~~1 for that feature.~~1. Lexicographic

311 features (length and first letter) are *discrete*. For these features, we defined a discrete  
 312 distance function. Specifically, we defined the distance between two words as either the  
 313 absolute difference between their lengths, or the absolute distance between their starting  
 314 letters in the English alphabet, respectively. For example, two words that started with the  
 315 same letter would have a “first letter” distance of 0, and a pair of words starting with ‘J’  
 316 and ‘A’ would have a first letter distance of 9. Because words’ lengths and letters’ positions  
 317 in the alphabet are always integers, these discrete distances always take on integer values.  
 318 Finally, the visual features (color and location) are *continuous* and *multivariate*, in that  
 319 each “feature” is defined by multiple (positive) real values. We defined the “color” and  
 320 “location” distances between two words as the Euclidean distances between their  $(r, g, b)$   
 321 color ~~or~~ vectors and  $(x, y)$  location vectors (specified as percentages of screen width and  
 322 height), respectively. Therefore, the color and location distance measures always take on  
 323 non-negative real values (upper-bounded at ~~441.67~~ 439.94 for color, or ~~27-in~~ 124.52 for  
 324 location, reflecting the distances between the corresponding maximally different vectors).

325 **Constructing feature-sorted lists.** Given a list of words, a relevant feature, and each  
 326 word’s value(s) for that feature, we developed a stochastic algorithm for (noisily) sorting  
 327 the words. The stochastic aspect of our sorting procedure enabled us to obtain unique  
 328 orderings for each participant. First, we choose a word uniformly and at random from  
 329 the set of words on the to-be-presented list. Second, we compute the distances between  
 330 the chosen word’s feature value(s) and the corresponding feature value(s) of all yet-to-be-  
 331 presented words. Third, we convert these distances (between the previously presented  
 332 word’s feature values,  $a$ , and ~~the candidate word’s~~ each of the  $W$  yet-to-be-presented  
 333 candidate words’ feature values,  $b_{i \in 1 \dots W}$ ) to similarity scores:



**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).

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

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

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

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

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



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

#### 354 **Adaptive condition**

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

360 All participants in the adaptive condition began the experiment by studying a set of  
361 four *initialization* lists. Words ~~and features~~ on these lists were presented in a randomized  
362 order (computed independently for each participant). These initialization lists were used  
363 to estimate each participant’s “memory fingerprint,” ~~defined~~ which we define below. At  
364 a high level, a participant’s memory fingerprint describes how they prioritize or con-  
365 sider different semantic, lexicographic, and/or visual features when they organize their  
366 memories.

367 Next, participants studied a sequence of 12 lists in three batches of four lists each. These  
368 batches came in three types: *random*, *stabilize*, and *destabilize*. The batch types determined

369 how words on the lists in that batch were ordered. Lists in each batch were always  
370 presented consecutively (e.g., a participant might receive four random lists, followed  
371 by four stabilize lists, followed by four destabilize lists). The batch orders were evenly  
372 counterbalanced across participants: there are six possible orderings of the three batches,  
373 and 10 participants were randomly assigned to each ordering sub-condition.

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

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

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

403 **Permutation-corrected feature clustering scores.** Suppose that two lists contain unequal  
404 numbers of items of each size. For example, suppose that list *A* contains all "large" items,  
405 whereas list *B* contains an equal mix of "large" and "small" items. For a participant  
406 recalling list *A*, any correctly recalled item will necessarily match the size of the previous  
407 correctly recalled item. In other words, successively recalling several list *A* items of the  
408 same size is essentially meaningless, since *any* correctly recalled list *A* word will be large.  
409 In contrast, successively recalling several list *B* items of the same size *could* be meaningful,  
410 since (early in the recall sequence) the yet-to-be-recalled items come from a mix of sizes.  
411 However, once all of the small items on list *B* have been recalled, the best possible next  
412 matching recall will be a large item. All subsequent correct recalls must also be large  
413 items—so for those later recalls it becomes difficult to determine whether the participant  
414 is successively recalling large items because they are organizing their memories according  
415 to size, or (alternatively), whether they are simply recalling the yet-to-be-recalled items  
416 in a random order. In general, the precise order and blend of feature values expressed  
417 in a given list, the order and number of correct recalls a participant makes, the number

418 of intervening presentation positions between successive recalls, and so on, can all affect  
419 the range of clustering scores that are possible to observe for a given list. An uncorrected  
420 clustering score therefore conflates participants' actual memory organization with other  
421 "nuisance" factors.

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

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

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

449 **Online “fingerprint” analysis.** The presentation orders of some lists in the adaptive con-  
450 dition of our experiment (see *Adaptive condition*) were sorted according to ~~participants’ each~~  
451 individual participant’s current memory fingerprint, estimated using all of the lists they  
452 had studied up to that point in the experiment. Because our experiment incorporated  
453 a speech-to-text component, all of the behavioral data for each participant could be an-  
454 alyzed just a few seconds after the conclusion of the recall intervals for each list. We  
455 used the Quail Python package (Heusser et al., 2017) to apply speech-to-text algorithms  
456 to the just-collected audio data, aggregate the data for the given participant, and estimate  
457 the participant’s memory fingerprint using all of their available data up to that point in  
458 the experiment. Two aspects of our implementation are worth noting: First, because  
459 memory fingerprints are computed independently for each list and then averaged across  
460 lists, the already-computed memory fingerprints for earlier lists could be cached and  
461 ~~loaded-retrieved~~ as needed in future computations. This meant that ~~our computations~~  
462 ~~pertaining to~~ updating our estimate of a participant’s memory fingerprint ~~only needed to~~  
463 ~~consider data from the~~ required computing feature and temporal clustering scores only for  
464 the single most recent list. Second, ~~each element of the null distributions of uncorrected~~  
465 ~~fingerprint scores (see Permutation-corrected feature clustering scores)~~ the clustering scores  
466 for each dimension of a participant’s memory fingerprint could be estimated indepen-  
467 dently from the others. ~~This enabled us to make use of the testing computers’ multi-core~~

468 CPU architectures by considering (in parallel) elements of the null distributions in batches  
 469 of eight (i.e., the number of CPU cores on each testing computer). Taken together, we were  
 470 able to, as could each element of the null distributions of uncorrected clustering scores  
 471 computed for each dimension (see *Permutation-corrected feature clustering scores*). This  
 472 enabled us to aggressively parallelize the fingerprint-updating procedure and compress  
 473 the relevant computations into just a few seconds of computing time. The combined pro-  
 474 cessing time for the speech-to-text algorithm, fingerprint computations, and permutation-  
 475 based ordering procedure (described next) easily fit within the inter-list intervals, where  
 476 participants paused for a self-paced break before moving on to study and recall the next  
 477 list.

478 **Ordering “stabilize” and “destabilize” lists by an estimated fingerprint.** In the adap-  
 479 tive condition of our experiment, the presentation orders for *stabilize* and *destabilize* lists  
 480 were chosen to either maximally or minimally (respectively) comport with participants’  
 481 memory fingerprints. Given a participant’s memory fingerprint and a to-be-presented set  
 482 of items, we designed a permutation-based procedure for ordering the items. First, we  
 483 dropped from the participant’s fingerprint the temporal clustering score. For the remain-  
 484 ing feature dimensions, we arranged the clustering scores in the fingerprint into a template  
 485 vector  $\vec{f}$ . Second, we computed  $n = 2500$  random permutations of the to-be-presented  
 486 items. These permutations served as candidate presentation orders. We sought to select  
 487 the specific order that most (or least) closely matched  $f$ . Third, for each random permu-  
 488 tation, we computed the (permutation-corrected) “fingerprint,” treating the permutation  
 489 as though it were a potential “perfect” recall sequence. (We did not include temporal  
 490 clustering scores in these fingerprints, since the temporal clustering score for every per-  
 491 mutation is always equal to 1.) This yielded a “simulated fingerprint” vector  $\vec{f}_p$  for each  
 492 permutation  $p$ . We used these simulated fingerprints to select a specific permutation  $i$

493 that either maximized (for stabilize lists) or minimized (for destabilize lists) the correlation  
494 between  $\hat{f}_i$  and  $f$ .

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

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

## 514 Factoring out the effects of temporal clustering

515 For a given list of words, if the values along two feature dimensions (e.g., category and size)  
516 are correlated, then the clustering scores for those two dimensions will also be correlated  
517 (Figs. S4, S9). When lists are sorted along a given feature dimension, the sorted feature  
518 values will also tend to be correlated with the serial positions of the words in the list.  
519 This means that the temporal clustering score will *also* tend to be correlated with the  
520 clustering scores for the sorted feature dimension. These correlations mean that it can be  
521 difficult to specifically identify when participants are using one feature versus another (or  
522 a manipulated feature versus temporal information) to organize or search their memories.

523

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

538 While these temporally corrected clustering scores are useful for identifying when  
539 feature clustering cannot be accounted for by temporal clustering alone, they are *not*



necessarily valid estimates of the “true” degree to which participants are organizing their memories along a given feature dimension. For example, on a list where the presentation order and feature values (along the given feature dimension) are perfectly correlated, the temporally corrected score will have an expected value of 0.5 no matter which words a participant recalls, or the order in which they recall them. Therefore these temporally corrected clustering scores are interpretable only to the extent that presentation order and feature value are decoupled.

## Analyses

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

Probability of first recall curves (Atkinson and Shiffrin, 1968; Postman and Phillips, 1965; Welch and Burnett, 1924) reflect the probability that an item will be recalled first, as a function of its serial position during encoding. We used an analogous approach to compute the proportion of trials on which each item (as a function of its presentation position) was recalled at each output position  $n$  (Hogan, 1975; Howard and Kahana, 1999; Polyn et al., 2009; Zhang et al., 2019). To carry out this analysis, we initialized (for each participant) a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then, for each list, we found the presentation index of the word that was recalled first, and we filled in that position in the matrix with a 1. Finally, we averaged over the rows of the matrix to obtain a 1 by 16 array of probabilities, for each participant. We used an analogous procedure to compute probability of  $n^{\text{th}}$  recall curves for each participant. Specifically, we filled in the corresponding matrices according to the  $n^{\text{th}}$  recall on each list that each participant made. When a given participant had made fewer than  $n$  recalls for a given list, we simply excluded that list from our analysis when computing that participant’s curve(s). The probability of first recall curve corresponds to a special case where  $n = 1$ .

564 We note that several other studies have used a slightly different approach to compute  
565 these curves, by correcting for the “availability” of a given word to be recalled. For  
566 example, if a participant recalls item 1, then item 2 on a given list, our approach places a  
567 0 into the item 1 column for that list when computing the “probability of second recall”  
568 curve. However, accounting for the fact that the participant had already recalled item  
569 1, an alternative approach (e.g., Farrell, 2010) would be to count the item 1 column as  
570 “unobserved” (i.e., missing data). Ultimately we chose to use the simpler variant of this  
571 approach in our work, but we direct the reader to further discussion of this issue in other  
572 work (Farrell, 2014; Moran and Goshen-Gottstein, 2014).

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

574 The lag-conditional response probability (lag-CRP) curve (Kahana, 1996) reflects the prob-  
575 ability of recalling a given item after the just-recalled item, as a function of ~~their~~ the items’  
576 relative encoding positions (lag). In other words, a lag of 1 indicates that a recalled item  
577 was presented immediately after the previously recalled item, and a lag of -3 indicates  
578 that a recalled item came three items before the previously recalled item. For each re-  
579 call transition (following the first recall), we computed the lag between the presentation  
580 positions of the just-recalled word ~~’s presentation position~~ and the next-recalled word ~~’s~~  
581 ~~presentation position. We~~. We then computed the proportions of transitions (between  
582 successively recalled words) for each lag, normalizing for the total numbers of possible  
583 transitions. In carrying out this analysis, we excluded all incorrect recalls and ~~successive~~  
584 repetitions (i.e., recalling ~~the same word twice in a row~~ a word that had already appeared  
585 in the current recall sequence). This yielded, for each list, a 1 by number-of-lags (-15 to  
586 +15; 30 lags in total, excluding lags of 0) array of conditional probabilities. We averaged  
587 these probabilities across lists to obtain a single lag-CRP for each participant. Because

588 transitions at large absolute lags are rare, these curves are typically displayed using range  
589 restrictions (Kahana, 2012).

## 590 **Serial position curve**

591 Serial position curves (Murdock, 1962) reflect the proportion of participants who remember  
592 each item as a function of the items' serial positions during encoding. For each participant,  
593 we initialized a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then,  
594 for each correct recall, we identified the presentation position of the word and entered a  
595 1 into that position (row: list; column: presentation position) in the matrix. This resulted  
596 in a matrix whose entries indicated whether or not the words presented at each position,  
597 on each list, were recalled by the participant (depending on whether the corresponding  
598 entries were set to 1 or 0). Finally, we averaged over the rows of the matrix to yield a  
599 1 by 16 array representing the proportion of words at each position that the ~~participant~~  
600 participants remembered.

## 601 **Identifying event boundaries**

602 We used the distances between feature values for successively presented words (see *Defin-*  
603 *ing feature-based distances*) to estimate "event boundaries" where the feature values changed  
604 more than usual (DuBrow and Davachi, 2016; Ezzyat and Davachi, 2011; Manning et al.,  
605 2016; Radvansky and Copeland, 2006; Swallow et al., 2011, 2009). For each list, for each  
606 feature dimension, we computed the distribution of distances between the feature values  
607 for successively presented words. We defined event boundaries (e.g., Fig. 3B) as occurring  
608 between any successive pair of words whose distances along the given feature dimension  
609 were greater than ~~one standard deviation~~ two standard deviations above the mean for  
610 that list. Note that, because event boundaries are defined for each feature dimension, each

individual list may contain several sets of event boundaries, each at different moments in the presentation sequence (depending on the feature dimension of interest).

## Transparency and openness

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

## Results

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

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	-0.290	126	-0.051	0.772	0.772	-2.387	1.768
Temp clust	10.632	126	1.882	< 0.001	< 0.001	7.786	14.386
Cat clust	10.148	126	1.796	< 0.001	< 0.001	7.324	13.778
Sz clust	12.033	126	2.129	< 0.001	< 0.001	9.030	15.918
Len clust	10.720	126	1.897	< 0.001	< 0.001	7.442	15.174
1 <sup>st</sup> ltr clust	6.679	126	1.182	< 0.001	< 0.001	4.490	9.611

**Table 1: Comparing memory in the feature-rich versus reduced conditions (all lists).** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all lists from each participant. Abbreviations used in this table are defined in Table S1.

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

To obtain a clean preliminary estimate of the consequences on memory of randomly  
varying the font colors and locations of presented words (versus holding the font color  
fixed at black, and ~~holding the display the words’~~ locations fixed at the center of the  
~~display~~screen), we compared participants’ performance on the ~~feature-rich~~feature-rich  
and *reduced* experimental conditions (see *Random conditions*, Fig. S1, ~~Tab. 1~~). In the  
~~feature-rich-condition~~feature-rich condition, the words’ colors and locations varied ran-  
domly~~across-words~~, and in the reduced condition, words were always presented in  
black, at the center of the display. Aggregating across all lists for each participant,  
we found no difference in recall accuracy (i.e., the proportions of ~~correctly-recalled~~  
~~words~~) ~~for feature-rich-words successfully recalled~~ for feature-rich versus reduced  
lists ( $t(126) = -0.290, p = 0.772$ ). However, participants in the ~~feature-rich~~feature-rich

condition clustered their recalls substantially more along every dimension we examined (temporal clustering:  $t(126) = 10.624, p < 0.001$ ; semantic category clustering:  $t(126) = 10.077, p < 0.001$ ; size clustering:  $t(126) = 11.829, p < 0.001$ ; word length clustering:  $t(126) = 10.639, p < 0.001$ ; first letter clustering:  $t(126) = 7.775, p < 0.001$ ; see *Permutation-corrected feature clustering scores* for more information about how we quantified each participant's clustering tendencies.) Taken together, these comparisons suggest that adding new features changes how participants organize their memories of studied words, even when those new features are independent of the words themselves and ~~even when the new features~~ vary randomly across words. We found no evidence that those additional uninformative features were distracting (in terms of their impact on memory performance), but they did affect participants' recall dynamics (measured via their clustering scores).

~~We also wondered whether adding these incidental visual features to later lists (after the participants had already studied impoverished lists), or removing the visual features from later lists (after the participants had already studied visually diverse lists) might affect memory performance.~~ A core assumption of our approach is that each participant organizes their memories in a unique way. We defined each participant's *memory fingerprint* as the set of their permutation-corrected clustering scores across all dimensions we tracked in our study, including their six feature-based clustering scores (category, size, length, first letter, color, and location) and their temporal clustering score. Conceptually, a participant's memory fingerprint describes their tendency to order, in their recall sequences (and presumably, organize in memory), the studied words along each dimension. If these memory fingerprints are truly unique to each participant, then we would expect that the estimated fingerprints computed for a given participant, on different lists, should be more similar than the estimated fingerprints computed for different participants. We reasoned that the feature-rich condition would provide the best opportunity to test this

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	4.553	66	0.233	< 0.001	< 0.001	2.427	7.262
Temp clust	2.268	66	0.181	0.027	0.053	0.437	4.425
Cat clust	3.684	66	0.220	< 0.001	0.001	1.733	5.732
Sz clust	1.629	66	0.100	0.108	0.162	-0.207	3.905
Len clust	-0.100	66	-0.010	0.921	0.921	-2.217	1.899
1 <sup>st</sup> ltr clust	-0.412	66	-0.045	0.681	0.818	-2.461	1.645

**Table 2: Comparing memory for early versus late lists in the feature-rich condition.** The *t*-tests reported in the table were carried out across-participants. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	2.434	60	0.134	0.018	0.027	0.493	4.910
Temp clust	0.986	60	0.061	0.328	0.328	-0.897	3.348
Cat clust	2.755	60	0.177	0.008	0.016	0.761	5.189
Sz clust	3.081	60	0.201	0.003	0.009	1.210	5.326
Len clust	3.762	60	0.261	< 0.001	0.002	1.604	6.821
1 <sup>st</sup> ltr clust	1.721	60	0.175	0.090	0.109	-0.138	4.098

**Table 3: Comparing memory for early versus late lists in the reduced condition.** The *t*-tests reported in the table were carried out across-participants. Abbreviations used in this table are defined in Table S1.

assumption, since the clustering scores would not be potentially confounded by order manipulations. To test our “unique memory fingerprint” assumption, we compared the similarity (correlation) between the fingerprint from a single list (from one participant) and (a) the average fingerprint from all other lists from the same participant versus (b) the average fingerprint from each other participant (across all of their lists). Repeating this procedure for all lists and participants, we found that participants’ fingerprints for a held-out list were reliably more similar to their fingerprints for other lists than they were to other participants’ fingerprints ( $t(70280) = 5.077$ ,  $p < 0.001$ ,  $d = 0.162$ ,  $CI = [3.086, 6.895]$ ). That within-participant fingerprint similarity (across lists) was greater than across-participant fingerprint similarity suggests that participants’ memory fingerprints are relatively stable across lists, and that each participant’s fingerprint is unique to them.

Beyond affecting participants’ memories for individual lists, we wondered how studying early lists (with versus without incidental visual features) might affect how participants remembered later lists (again, with versus without incidental visual features). In other

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	1.499	41	0.098	0.141	0.580	-0.345	3.579
Temp clust	0.857	41	0.068	0.396	0.580	-1.012	2.896
Cat clust	0.707	41	0.068	0.484	0.580	-1.314	2.830
Sz clust	0.803	41	0.079	0.427	0.580	-1.142	2.953
Len clust	0.461	41	0.060	0.648	0.648	-1.545	2.462
1 <sup>st</sup> ltr clust	0.781	41	0.101	0.439	0.580	-1.039	2.881

**Table 4:** Comparing memory for early versus late lists in the reduced (early) condition. The *t*-tests reported in the table were carried out across-participants. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	1.462	40	0.121	0.152	0.441	-0.376	2.993
Temp clust	1.244	40	0.128	0.221	0.441	-0.894	3.088
Cat clust	-0.101	40	-0.009	0.920	0.920	-2.307	1.776
Sz clust	0.555	40	0.058	0.582	0.873	-1.444	2.274
Len clust	1.482	40	0.126	0.146	0.441	-0.444	3.743
1 <sup>st</sup> ltr clust	-0.143	40	-0.017	0.887	0.920	-2.204	1.830

**Table 5:** Comparing memory for early versus late lists in the reduced (late) condition. The *t*-tests reported in the table were carried out across-participants. Abbreviations used in this table are defined in Table S1.

words, we sought to test for potential effects of changing the “richness” of participants’ experiences over time. All participants studied and recalled a total of 16 lists; we defined *early* lists as the first eight lists and *late* lists as the last eight lists each participant encountered. To help interpret our results, we compared participants’ memories on early versus late lists in the above ~~feature-rich and reduced~~ feature-rich (Tab. 2) and reduced (Tab. 3) conditions. Participants in both conditions remembered more words on early versus late lists (~~feature-rich:  $t(66) = 4.553, p < 0.001$ ; reduced:  $t(60) = 2.434, p = 0.018$~~ ). Participants in the ~~feature-rich~~ feature-rich (but not reduced) ~~conditions~~ condition exhibited more temporal clustering on early versus late lists (~~feature-rich:  $t(66) = 2.318, p = 0.024$ ; reduced:  $t(60) = 0.929, p = 0.357$~~ ). And participants in both conditions ~~exhibited more semantic (category and size)~~ tended to exhibit more semantic clustering on early versus late lists (~~feature-rich, category:  $t(66) = 3.805, p < 0.001$ ; feature-rich, size:  $t(66) = 2.190, p = 0.032$ ; reduced, category:  $t(60) = 2.856, p = 0.006$ ; reduced, size:  $t(60) = 2.947, p = 0.005$~~ ). Participants in the reduced (but not ~~feature-rich~~) ~~conditions~~ exhibited feature-rich conditions



tended to exhibit more lexicographic clustering on early versus late lists (feature-rich, word length:  $t(66) = 0.161, p = 0.872$ ; feature-rich, first letter:  $t(66) = 0.410, p = 0.683$ ; reduced, word length:  $t(60) = 3.528, p = 0.001$ ; reduced, first letter:  $t(60) = 2.275, p = 0.026$ ). Taken together, these comparisons suggest that even when the presence or absence of incidental visual features is stable across lists, participants still exhibit some differences in their performance and memory organization tendencies for early versus late lists.

With these differences in mind, we next compared participants' memories on early versus late lists for two additional experimental conditions (see *Random conditions*, Fig. S1). In a *reduced (early)* condition, we held the visual features constant on early lists, but allowed them to vary randomly on late lists. In a *reduced (late)* condition, we allowed the visual features to vary randomly on early lists, but held them constant on late lists. Given our above findings that (a) participants tended to remember more words and exhibit stronger clustering effects on feature-rich feature-rich (versus reduced) lists, and (b) participants tended to remember more words and exhibit stronger clustering effects on early (versus late) lists, we expected these early versus late differences to be enhanced in the reduced (early) condition and diminished in the reduced (late) condition. However, to our surprise, participants in neither condition exhibited reliable early-versus-late early-versus-late differences in accuracy (reduced (early):  $t(41) = 1.499, p = 0.141$ ; reduced (late):  $t(40) = 1.462, p = 0.152$ ), temporal clustering (reduced (early):  $t(41) = 0.998, p = 0.324$ ; reduced (late):  $t(40) = 1.099, p = 0.278$ ), nor feature-based clustering (reduced (early), category:  $t(41) = 0.753, p = 0.456$ ; reduced (early), size:  $t(41) = 0.721, p = 0.475$ ; reduced (early), length:  $t(41) = 0.493, p = 0.625$ ; reduced (early), first letter:  $t(41) = 0.780, p = 0.440$ ; reduced (late), category:  $t(40) = -0.086, p = 0.932$ ; reduced (late), size:  $t(40) = 0.746, p = 0.460$ ; reduced (late), length:  $t(40) = 1.476, p = 0.148$ ; reduced (late), first letter:  $t(40) = 0.966, p = 0.340$ ) (Tabs. 4, 5). We hypothesized that adding or removing the variability in the visual features was acting as a sort of "event boundary"

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	-2.230	107	-0.439	0.028	0.167	-4.252	-0.229
Temp clust	-1.379	107	-0.271	0.171	0.512	-3.319	0.474
Cat clust	0.013	107	0.003	0.989	0.989	-2.003	2.102
Sz clust	-0.349	107	-0.069	0.728	0.873	-2.244	1.641
Len clust	-0.581	107	-0.114	0.563	0.844	-2.328	1.291
1 <sup>st</sup> ltr clust	0.636	107	0.125	0.526	0.844	-1.291	2.940

**Table 6: Comparing memory in the feature-rich versus reduced (early) conditions (all lists).** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all lists from each participant. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	-2.045	101	-0.410	0.043	0.043	-3.826	0.112
Temp clust	-10.689	101	-2.143	< 0.001	< 0.001	-13.479	-8.512
Cat clust	-9.538	101	-1.912	< 0.001	< 0.001	-12.332	-7.457
Sz clust	-12.222	101	-2.451	< 0.001	< 0.001	-15.311	-9.954
Len clust	-10.620	101	-2.129	< 0.001	< 0.001	-13.902	-8.239
1 <sup>st</sup> ltr clust	-5.213	101	-1.045	< 0.001	< 0.001	-7.290	-3.403

**Table 7: Comparing memory in the reduced versus reduced (early) conditions (all lists).** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all lists from each participant. Abbreviations used in this table are defined in Table S1.

between early and late lists (e.g., Clewett et al., 2019; Radvansky and Copeland, 2006; Radvansky and Zacks, 2016). In prior work, we (and others) have found that memories formed just after event boundaries can be enhanced (e.g., due to less contextual interference between pre- and post-boundary items; Flores et al., 2017; Gold et al., 2017; Manning et al., 2016; Pettijohn et al., 2016).

We found that *adding* incidental visual features on later lists that had not been present on early lists (as in the reduced (early) condition) served to enhance recall performance relative to conditions where all lists had the same blends of features (accuracy for feature rich versus reduced (early):  $t(107) = -2.230, p = 0.028$ ; reduced versus reduced (early):  $t(101) = -2.045, p = 0.043$ ; Tabs. 6, 7; also see Fig. S3A). However, *subtracting* irrelevant visual features on later lists that *had* been present on early lists (as in the reduced (late) condition) did not appear to impact recall performance (accuracy for feature rich versus reduced (late):  $t(106) = -0.638, p = 0.525$ ; reduced versus reduced (late):  $t(100) = -0.407, p = 0.685$ ; Tabs. 8, 9). These comparisons suggest that recall accuracy has a directional component: accuracy is

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	-0.638	106	-0.126	0.525	0.593	-2.720	1.362
Temp clust	-0.535	106	-0.106	0.593	0.593	-2.552	1.237
Cat clust	-1.345	106	-0.267	0.181	0.420	-3.525	0.660
Sz clust	-1.441	106	-0.286	0.153	0.420	-3.557	0.382
Len clust	-1.261	106	-0.250	0.210	0.420	-3.611	0.669
1 <sup>st</sup> ltr clust	0.939	106	0.186	0.350	0.525	-1.018	2.949

**Table 8:** Comparing memory in the feature-rich versus reduced (late) conditions (all lists). The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all lists from each participant. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Acc	-0.407	100	-0.082	0.685	0.685	-2.477	1.626
Temp clust	-9.885	100	-1.996	< 0.001	< 0.001	-14.701	-6.499
Cat clust	-10.436	100	-2.107	< 0.001	< 0.001	-15.607	-6.940
Sz clust	-12.413	100	-2.507	< 0.001	< 0.001	-18.413	-8.398
Len clust	-9.672	100	-1.953	< 0.001	< 0.001	-14.476	-6.437
1 <sup>st</sup> ltr clust	-4.555	100	-0.920	< 0.001	< 0.001	-7.332	-2.538

**Table 9:** Comparing memory in the reduced versus reduced (late) conditions (all lists). The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all lists from each participant. Abbreviations used in this table are defined in Table S1.

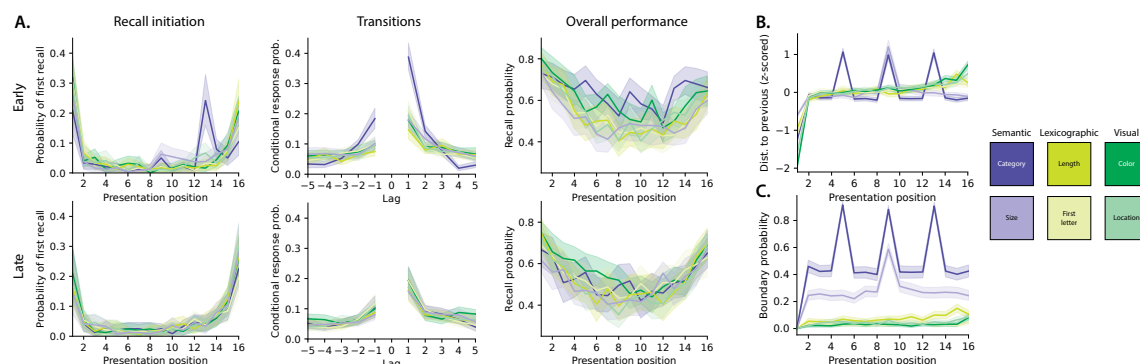
affected differently by removing features ~~later that had been present earlier that had~~  
initially been present versus adding features ~~later that had not been present earlier that~~  
had initially been absent. In contrast, we found that participants exhibited more tem-  
poral and feature-based clustering when we added incidental visual features to *any* lists  
(comparisons of clustering on feature rich versus reduced lists are reported above; temporal  
clustering in feature-rich versus reduced: Tab. 1; reduced versus reduced (early) and  
reduced versus reduced (late) conditions:  $ts \leq -9.780$ ,  $ps < 0.001$ ; feature-based clustering  
in: Tab. 7; reduced versus reduced (early) and reduced versus reduced (late) conditions:  
 $ts \leq -5.443$ ,  $ps < 0.001$  late): Tab. 9). Temporal and feature-based clustering were not reli-  
ably different in the ~~feature rich, reduced (early), and reduced (late) conditions (temporal~~  
~~clustering in feature rich feature-rich~~ versus reduced (early) ~~and feature rich versus or~~  
reduced (late) conditions :  $ts \geq -1.434$ ,  $ps \geq 0.154$ ; feature-based clustering in feature rich  
versus reduced (early) and feature rich versus reduced (late) conditions:  $ts \geq -1.359$ ,  $ps$   
 $\geq 0.177$  (Tabs. 6, 8).

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	3.034	95	0.667	0.003	0.019	1.048	5.113
Sz	-1.013	95	-0.223	0.314	0.627	-3.055	0.865
Len	-0.550	95	-0.121	0.584	0.700	-2.368	1.363
1 <sup>st</sup> ltr	-0.690	95	-0.152	0.492	0.700	-2.663	1.119
Clr	1.850	96	0.402	0.067	0.202	-0.010	3.712
Loc	0.043	95	0.010	0.966	0.966	-1.598	1.729

**Table 10: Comparing accuracy on early lists in the order manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. Abbreviations used in this table are defined in Table S1.

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

Across each of six order manipulation conditions, we sorted early lists by one feature dimension but randomly ordered the items on late lists (see *Order manipulation conditions*;



**Figure 3: Recall dynamics in feature rich free recall (order manipulation conditions).** 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.** Distances between successively presented words (z-scored within condition) computed based on each condition's feature of focus, and plotted as a function of presentation position. See *Defining feature-based distances* for additional information. **C.** Proportion of event boundaries (see *Identifying event boundaries*) for each condition's feature of focus, plotted as a function of presentation position. **All panels.** Error ribbons denote bootstrap-estimated 95% confidence intervals (calculated across participants in Panel A, and across lists in Panels B and C).

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	8.813	95	1.936	< 0.001	< 0.001	6.793	11.751
Sz	2.630	95	0.578	0.010	0.020	0.831	4.866
Len	-1.547	95	-0.340	0.125	0.150	-3.693	0.341
1 <sup>st</sup> ltr	2.858	95	0.628	0.005	0.016	1.031	4.886
Clr	-1.339	96	-0.291	0.184	0.184	-3.238	0.394
Loc	1.705	95	0.374	0.092	0.137	-0.155	3.521

**Table 11: Comparing temporal clustering on early lists in the order manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	2.722	125	0.484	0.007	0.022	0.827	4.932
Sz	3.866	125	0.687	< 0.001	0.001	2.020	5.983
Len	0.521	125	0.093	0.603	0.724	-1.311	2.333
1 <sup>st</sup> ltr	-0.842	125	-0.150	0.401	0.724	-2.825	1.095
Clr	-0.650	125	-0.116	0.517	0.724	-2.680	1.249
Loc	-0.251	125	-0.045	0.802	0.802	-2.257	1.524

**Table 12: Comparing feature-based clustering on early lists in the semantic order manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	-1.040	125	-0.185	0.301	0.601	-3.095	1.092
Sz	0.006	125	0.001	0.995	0.995	-1.933	1.952
Len	3.682	125	0.655	< 0.001	0.001	1.890	5.569
1 <sup>st</sup> ltr	5.134	125	0.912	< 0.001	< 0.001	3.251	7.258
Clr	0.092	125	0.016	0.927	0.995	-1.834	1.867
Loc	0.407	125	0.072	0.685	0.995	-1.655	2.463

**Table 13: Comparing feature-based clustering on early lists in the lexicographic order manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. Abbreviations used in this table are defined in Table S1.

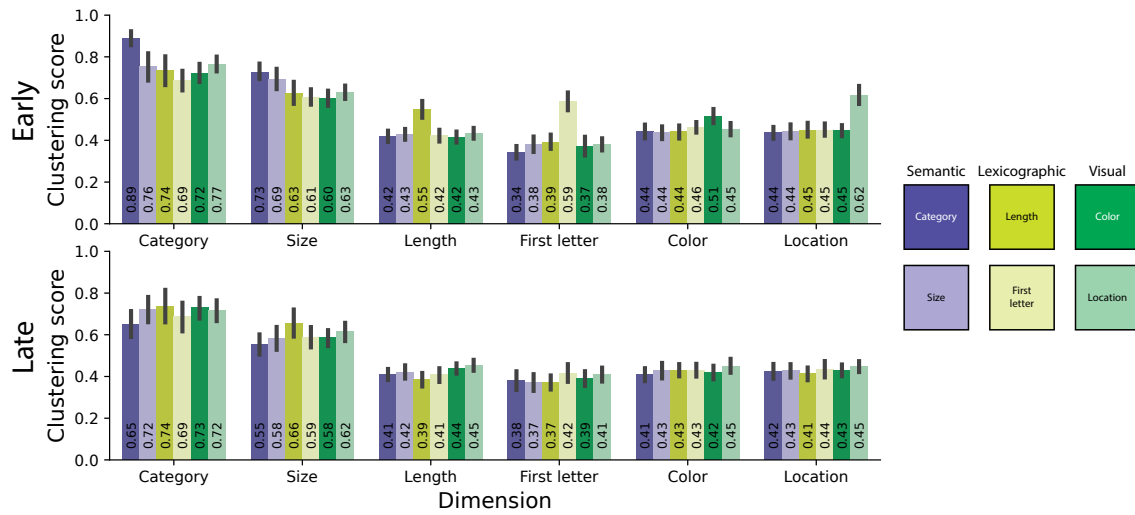
	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	0.012	126	0.002	0.991	0.991	-1.988	1.871
Sz	-0.104	126	-0.018	0.917	0.991	-2.166	1.847
Len	0.592	126	0.105	0.555	0.991	-1.361	2.420
1 <sup>st</sup> ltr	0.040	126	0.007	0.968	0.991	-1.791	1.863
Clr	2.022	126	0.358	0.045	0.136	0.056	3.965
Loc	4.390	126	0.777	< 0.001	< 0.001	2.730	6.199

**Table 14: Comparing feature-based clustering on early lists in the visual order manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. Abbreviations used in this table are defined in Table S1.

778 features: category, size, length, first letter, color, and location). ~~Participants in the~~ When  
 779 ~~we compared participants' memories for early lists in each of these conditions to their~~  
 780 ~~memories for early lists in the feature-rich condition (Tab. 10), we found that participants~~  
 781 ~~in the~~ category-ordered ~~condition showed an increase in memory performance on early~~  
 782 ~~lists (accuracy, relative to early feature rich lists ;  $t(95) = 3.034, p = 0.003$ ).~~ Participants  
 783 ~~in the color-ordered condition also showed a trending increase in memory performance~~  
 784 ~~on early lists (again, relative to early feature rich lists:  $t(96) = 1.850, p = 0.067$ )~~ condition  
 785 ~~remembered more words than participants in the feature-rich condition.~~ Participants'  
 786 performances on early lists in all of the other order manipulation conditions were in-  
 787 distinguishable from performance on the early ~~feature rich lists ( $|t|s < 1.013, ps > 0.314$ ).~~  
 788 ~~Participants in both of the semantically ordered conditions exhibited stronger~~ feature-rich  
 789 ~~lists. We also compared participants' temporal clustering on early lists (versus early feature~~  
 790 ~~rich lists; category:  $t(95) = 8.508, p < 0.001$ ; size:  $t(95) = 2.429, p = 0.017$ ).~~ Participants in  
 791 ~~the length-ordered condition tended to exhibit less~~ in each of these conditions to their tem-  
 792 poral clustering on early lists ~~relative to early feature rich lists ( $t(95) = -1.666, p = 0.099$ ),~~  
 793 ~~whereas participants in the~~ in the feature-rich condition (Tab. 11). Participants in both  
 794 ~~of the semantically ordered conditions and the~~ first letter-ordered condition all exhibited  
 795 stronger temporal clustering on early lists ( ~~$t(95) = 2.587, p = 0.011$~~  vs. early feature-rich  
 796 ~~lists~~). Participants in the ~~visually ordered conditions exhibited more similar performance~~  
 797 ~~other order manipulation conditions all showed similar temporal clustering~~ on early lists  
 798 ~~, relative to early feature rich lists (color:  $t(96) = -1.064, p = 0.290$ ; we found a trending~~  
 799 ~~enhancement for participants in the location-ordered condition:  $t(95) = 1.682, p = 0.096$ )~~ feature-rich  
 800 ~~lists.~~ We also ~~also~~ compared feature-based clustering on early lists across the order  
 801 manipulation and ~~feature-rich~~ feature-rich conditions. Since these results were similar  
 802 across both semantic conditions (category and size; Tab. 12), both lexicographic condi-

tions (length and first letter; [Tab. 13](#)), and both visual conditions (color and location; [Tab. 14](#)), here we aggregate data from conditions that manipulated each of these three feature groupings in our comparisons, to simplify the presentation. On early lists, participants in the semantically ordered conditions exhibited stronger semantic clustering relative to participants in the ~~feature rich condition~~(category:  $t(125) = 2.524, p = 0.013$ ; size:  $t(125) = 3.510, p = 0.001$ )[feature-rich condition](#), but showed no reliable differences in lexicographic (~~length:  $t(125) = 0.539, p = 0.591$ ; first letter:  $t(125) = -0.587, p = 0.558$~~ ) or visual (~~color:  $t(125) = -0.579, p = 0.564$ ; location:  $t(125) = -0.346, p = 0.730$~~ ) or [visual](#) clustering. Similarly, participants in the lexicographically ordered conditions exhibited stronger (relative to feature rich participants) lexicographic clustering (~~length:  $t(125) = 3.426, p = 0.001$ ; first letter:  $t(125) = 3.236, p = 0.002$~~ ) on early lists, but showed no reliable differences in semantic (~~category:  $t(125) = -1.078, p = 0.283$ ; size:  $t(125) = -0.310, p = 0.757$~~ ) or visual (~~color:  $t(125) = -0.209, p = 0.835$ ; location:  $t(125) = -0.004, p = 0.997$~~ ) or [visual](#) clustering. And participants in the visually ordered conditions exhibited stronger visual clustering (again, relative to ~~feature rich~~[feature-rich](#) participants, and on early lists; ~~color:  $t(126) = 2.099, p = 0.038$ ; location:  $t(126) = 4.392, p < 0.001$~~ ), but showed no reliable differences in semantic (~~category:  $t(126) = 0.204, p = 0.839$ ; size:  $t(126) = -0.093, p = 0.926$~~ ) or lexicographic (~~length:  $t(126) = 0.714, p = 0.476$ ; first letter:  $t(126) = 0.820, p = 0.414$~~ ) or [lexicographic](#) clustering. Taken together, these order manipulation results suggest several broad patterns (Figs. 3A, 4). First, most of the order manipulations we carried out did *not* reliably affect overall recall performance. Second, most of the order manipulations increased participants' tendencies to temporally cluster their recalls. Third, all of the order manipulations enhanced participants' clustering of each condition's target feature (i.e., semantic manipulations enhanced semantic clustering, lexicographic manipulations enhanced lexicographic clustering, and visual manipulations enhanced visual cluster-





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

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

When we closely examined the sequences of words participants recalled from early  
 order-manipulated lists (Fig. 3A, top panel), we noticed several differences from the dy-

namics of participants' recalls of randomly ordered lists (Figs. S1, S7). One difference  
 is that participants in the category condition (~~dark purple curves~~, Fig. 3, dark purple  
curves) most often initiated recall with the fourth-from-last item (*Recall initiation*, top left  
 panel), whereas participants who recalled randomly ordered lists tended to initiate recall  
 with either the first or last list items (Fig. S1, top left panel). We hypothesized that the  
 participants might be "clumping" their recalls into groups of items that shared category  
 labels. Indeed, when we compared the positions of feature changes in the study sequence  
 (Fig. 3BC; see *Identifying event boundaries*) with the positions of items participants recalled  
 first, we noticed a striking correspondence in both semantic conditions. Specifically, on  
 category-ordered lists, the category labels changed every four items on average (dark pur-  
 ple peaks in ~~FigFigs.~~ 3B, C), and participants also seemed to display an increased tendency  
 (relative to other order manipulation and random conditions) to initiate recall of category-  
 ordered lists with items whose study positions were integer multiples of four. Similarly,  
 for size-ordered lists, the size labels changed every eight items on average (light purple  
 peaks in ~~FigFigs.~~ 3B, C), and participants also seemed to display an increased tendency  
 to initiate recall of size-ordered lists with items whose study positions were integer mul-  
 tiples of eight. A second striking difference is that participants in the category condition  
 exhibited a much steeper lag-CRP (Fig. 3A, top middle panel) than participants in other  
 conditions. (This is another expression of participants' increased tendencies to tempo-  
 rally cluster their recalls on category-ordered lists, as we reported above.) Taken together,  
 these order-specific idiosyncrasies suggest a hierarchical set of influences on participants'  
 memories. At longer timescales, "event boundaries" (to use the term loosely) can be in-  
 duced across lists by adding or removing incidental visual features. At shorter timescales,  
 "event boundaries" can be induced across items (within a single list) by adjusting how  
 item features change throughout the list.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Sem vs. lex	1.936	118	0.353	0.055	0.083	0.057	3.916
Sem vs. vis	0.113	119	0.021	0.910	0.910	-1.987	2.097
Lex vs. vis	-2.145	119	-0.390	0.034	0.083	-4.254	-0.208

**Table 15: Comparing accuracy on early lists in different order manipulation conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Sem vs. lex	5.620	118	1.026	< 0.001	< 0.001	3.486	8.010
Sem vs. vis	6.613	119	1.202	< 0.001	< 0.001	4.481	9.464
Lex vs. vis	0.589	119	0.107	0.557	0.557	-1.336	2.539

**Table 16: Comparing temporal clustering on early lists in different order manipulation conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat: sem vs. lex	3.667	118	0.670	< 0.001	< 0.001	1.822	5.942
Sz: sem vs. lex	4.043	118	0.738	< 0.001	< 0.001	2.145	6.296
Len: sem vs. lex	-3.390	118	-0.619	< 0.001	0.002	-5.661	-1.499
1 <sup>st</sup> ltr: sem vs. lex	-5.705	118	-1.042	< 0.001	< 0.001	-7.790	-3.841
Clr: sem vs. lex	-0.767	118	-0.140	0.444	0.533	-2.744	1.154
Loc: sem vs. lex	-0.658	118	-0.120	0.512	0.576	-2.595	1.171
Cat: sem vs. vis	3.114	119	0.566	0.002	0.004	1.052	5.737
Sz: sem vs. vis	4.692	119	0.853	< 0.001	< 0.001	2.620	7.024
Len: sem vs. vis	-0.068	119	-0.012	0.946	0.946	-1.897	1.907
1 <sup>st</sup> ltr: sem vs. vis	-0.842	119	-0.153	0.401	0.516	-2.944	1.089
Clr: sem vs. vis	-2.673	119	-0.486	0.009	0.014	-4.567	-0.848
Loc: sem vs. vis	-4.499	119	-0.818	< 0.001	< 0.001	-6.399	-2.721
Cat: lex vs. vis	-1.186	119	-0.216	0.238	0.329	-3.010	0.891
Sz: lex vs. vis	0.118	119	0.021	0.906	0.946	-1.778	2.271
Len: lex vs. vis	3.399	119	0.618	< 0.001	0.002	1.500	5.527
1 <sup>st</sup> ltr: lex vs. vis	4.859	119	0.883	< 0.001	< 0.001	2.860	6.849
Clr: lex vs. vis	-1.988	119	-0.361	0.049	0.074	-3.894	-0.102
Loc: lex vs. vis	-3.966	119	-0.721	< 0.001	< 0.001	-5.862	-2.099

**Table 17: Comparing feature-based clustering on early lists in different order manipulation conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all early lists from each participant. The feature used to compute clustering is shown before the colon in each row, and the conditions being compared are shown after the colon. Abbreviations used in this table are defined in Table S1.

865 The above comparisons between memory performance on early lists in the order  
 866 manipulation ~~versus feature-rich~~ and feature-rich conditions highlight how sorted lists  
 867 are remembered differently from random lists. We also wondered how sorting lists  
 868 along each feature dimension influenced memory relative to sorting lists along the other  
 869 feature dimensions ~~—(accuracy: Tab. 15; temporal clustering: Tab. 16; feature-based~~  
 870 clustering: Tab 17). Participants trended towards remembering early lists that were  
 871 sorted ~~semantically better than lexicographically sorted lists ( $t(118) = 1.936, p = 0.055$ ).~~  
 872 ~~Participants also remembered visually sorted lists~~ visually better than lexicographically  
 873 sorted lists ( ~~$t(119) = 2.145, p = 0.034$~~ ). However, participants showed no reliable differ-  
 874 ences in recall for semantically versus lexicographically or visually sorted lists ( ~~$t(119) = 0.113, p = 0.910$~~ ).  
 875 Participants temporally clustered semantically sorted lists more strongly than ~~either~~  
 876 ~~lexicographically ( $t(118) = 5.572, p < 0.001$ ) or visually ( $t(119) = 6.215, p < 0.001$ ) sorted lists~~ lists  
 877 sorted either lexicographically or visually, but did not show reliable differences in tem-  
 878 poral clustering on lexicographically versus visually sorted lists ( ~~$t(119) = 0.189, p = 0.850$~~ ).  
 879 Participants also showed reliably more semantic clustering on semantically sorted lists  
 880 than lexicographically (~~category:  $t(118) = 3.492, p = 0.001$ , size:  $t(118) = 3.972, p < 0.001$~~   
 881 ~~or visually (category:  $t(119) = 2.702, p = 0.008$ , size:  $t(119) = 4.230, p < 0.001$ ) or visually~~  
 882 sorted lists; more lexicographic clustering on lexicographically sorted lists than seman-  
 883 tically (~~length:  $t(118) = 3.112, p = 0.002$ ; first letter:  $t(118) = 3.686, p < 0.001$ ) or visually~~  
 884 (~~length:  $t(119) = 3.024, p = 0.003$ ; first letter:  $t(119) = 2.644, p = 0.009$ ) or visually~~ sorted  
 885 lists; and more visual clustering on visually sorted lists than semantically (~~color:  $t(119) = -2.659, p = 0.009$ ;~~  
 886 ~~location:  $t(119) = -4.604, p < 0.001$ ) or lexicographically (color:  $t(119) = -2.366, p = 0.020$ ;~~  
 887 ~~location:  $t(119) = -4.265, p < 0.001$ ) or lexicographically~~ sorted lists. In summary, sorting  
 888 lists by different features appeared to have slightly different effects on overall memory per-  
 889 formance and temporal clustering. Participants also tended to cluster their recalls along a

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Sem	0.487	125	0.087	0.627	0.627	-1.661	2.323
Lex	0.878	125	0.156	0.382	0.573	-1.226	3.044
Vis	1.437	126	0.254	0.153	0.460	-0.447	3.519

**Table 18: Comparing accuracy on late lists in order-manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all late lists from each participant. Abbreviations used in this table are defined in Table S1.

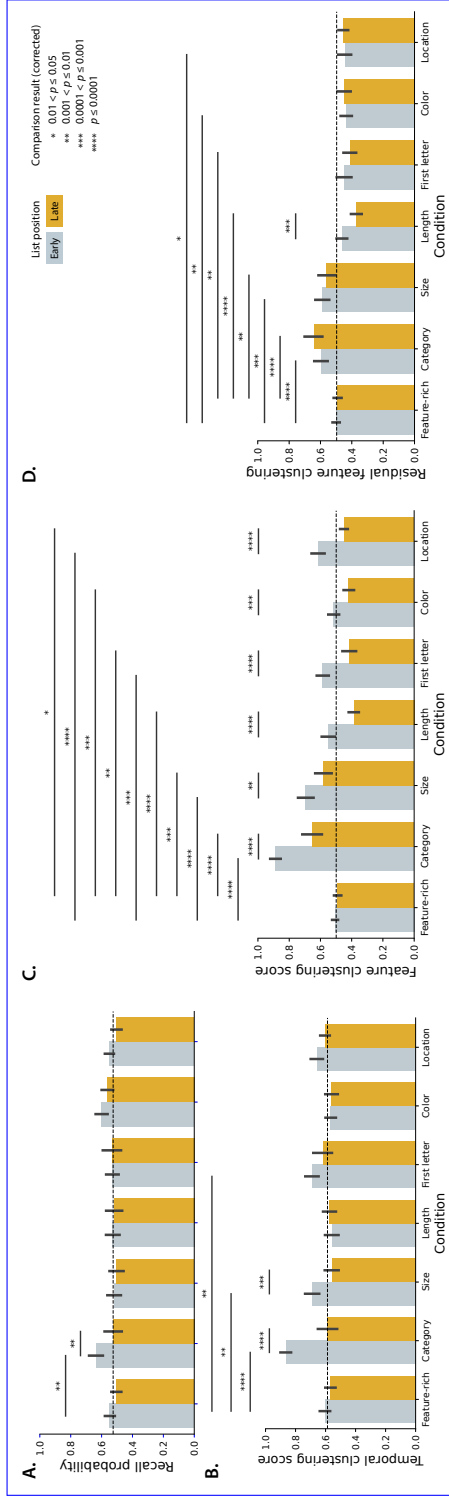
	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Sem	0.157	125	0.028	0.875	0.875	-1.859	1.974
Lex	0.998	125	0.177	0.320	0.875	-0.902	2.920
Vis	0.548	126	0.097	0.585	0.875	-1.450	2.365

**Table 19: Comparing temporal clustering on late lists in order-manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all late lists from each participant. Abbreviations used in this table are defined in Table S1.

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

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

At the group level, there seemed to be almost no lingering impact of sorting early lists on memory for later lists. To simplify the presentation, we report these null results in aggregate across the three feature groupings —(accuracy: Tab. 18; temporal clustering: Tab. 19; feature-based clustering: Tabs. 20, 21, and 22). Relative to memory performance on late ~~feature-rich~~ feature-rich lists, participants' memory performance in all six order manipulation conditions showed no reliable differences(semantic:  $t(125) = 0.487, p = 0.627$ ;



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

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	-0.041	125	-0.007	0.967	0.967	-2.088	1.861
Sz	-0.989	125	-0.176	0.324	0.967	-3.100	0.948
Len	-0.045	125	-0.008	0.964	0.967	-1.959	1.870
1 <sup>st</sup> ltr	-0.369	125	-0.066	0.713	0.967	-2.338	1.630
Clr	-0.602	125	-0.107	0.548	0.967	-2.541	1.273
Loc	-0.521	125	-0.093	0.603	0.967	-2.592	1.565

**Table 20: Comparing feature-based clustering on late lists in semantic order-manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all late lists from each participant. Abbreviations used in this table are defined in Table S1.

	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	0.678	125	0.121	0.499	0.655	-1.240	2.608
Sz	0.915	125	0.163	0.362	0.655	-1.137	2.756
Len	-1.200	125	-0.213	0.233	0.655	-3.499	0.737
1 <sup>st</sup> ltr	0.606	125	0.108	0.546	0.655	-1.390	2.553
Clr	0.094	125	0.017	0.925	0.925	-1.955	1.966
Loc	-0.619	125	-0.110	0.537	0.655	-2.672	1.270

**Table 21: Comparing feature-based clustering on late lists in lexicographic order-manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all late lists from each participant. Abbreviations used in this table are defined in Table S1.

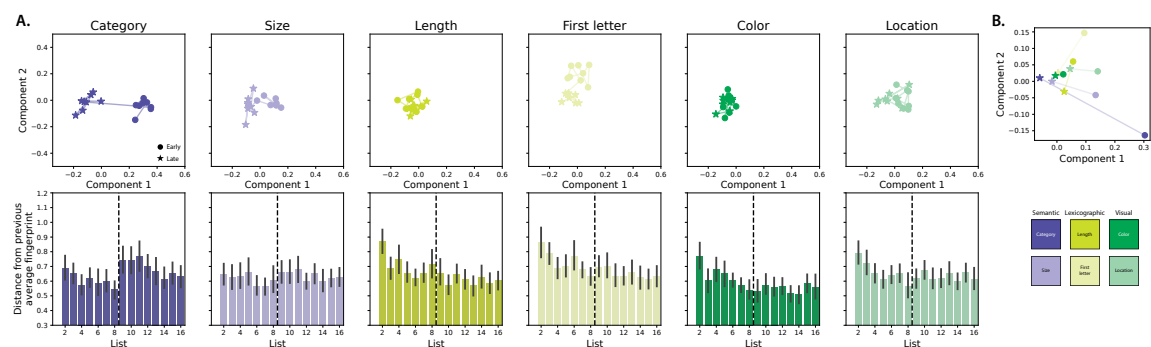
	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Cat	1.209	126	0.214	0.229	0.526	-0.700	3.136
Sz	0.202	126	0.036	0.840	0.869	-1.832	2.163
Len	2.005	126	0.355	0.047	0.283	0.211	3.722
1 <sup>st</sup> ltr	1.124	126	0.199	0.263	0.526	-0.846	3.260
Clr	0.278	126	0.049	0.781	0.869	-1.710	2.084
Loc	0.165	126	0.029	0.869	0.869	-1.779	2.004

**Table 22: Comparing feature-based clustering on late lists in visual order-manipulation versus feature-rich conditions.** The *t*-tests reported in the table were carried out across-participants, and reflect data aggregated across all late lists from each participant. Abbreviations used in this table are defined in Table S1.

905 ~~lexicographic:  $t(125) = 0.878, p = 0.382$ ; visual:  $t(126) = 1.437, p = 0.153$~~ . Nor did we ob-  
 906 serve any reliable differences in temporal clustering on late lists (relative to late ~~feature-rich~~  
 907 ~~lists; semantic:  $t(125) = 0.146, p = 0.884$ ; lexicographic:  $t(125) = 0.923, p = 0.358$ ; visual:~~  
 908  ~~$t(126) = 0.525, p = 0.601$~~ feature-rich lists). Aside from a slightly increased tendency for  
 909 participants to cluster words by their length on late visual order manipulation lists (more  
 910 than late ~~feature-rich lists;  $t(126) = 2.199, p = 0.030$~~ feature-rich lists), we observed no reli-  
 911 able differences in any type of feature clustering on late order manipulation condition lists  
 912 versus late ~~feature-rich lists ( $||t||s \leq 1.234, ps \geq 0.220$ )~~feature-rich lists.

913 We also looked for more subtle group-level patterns. For example, perhaps sorting  
 914 early lists by one feature dimension could affect how participants cluster *other* features  
 915 (on early and/or late lists) as well. ~~We defined participants' memory fingerprints as the set~~  
 916 ~~of their temporal and feature clustering scores (see Memory fingerprints).~~ ~~A~~ As described  
 917 above, a participant's memory fingerprint ~~describes~~ characterizes how they tend to retrieve  
 918 memories of the studied items, perhaps searching in parallel through several feature  
 919 spaces (or along several representational dimensions). To gain insights into the dynamics  
 920 of how participants' clustering scores tended to change over time, we computed the  
 921 average (across participants) fingerprint from each list, from each order manipulation  
 922 condition (Fig. 6). We projected these fingerprints into a two-dimensional space to help  
 923 visualize the dynamics (top panels; see *Computing low-dimensional embeddings of memory*  
 924 *fingerprints*). We found that participants' average fingerprints tended to remain relatively  
 925 stable on early lists, and exhibited a "jump" to another stable state on later lists. The  
 926 sizes of these jumps varied somewhat across conditions (the Euclidean distances between  
 927 fingerprints in their original high dimensional spaces are displayed in the bottom panels).  
 928 We also averaged the fingerprints across early and late lists, respectively, for each condition  
 929 (Fig. 6B). We found that participants' fingerprints on early lists seem to be influenced by





**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 a single list. Lines connect successive lists. 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 between each per-list memory fingerprints to fingerprint and the list-0 average fingerprint across all prior lists, for each condition. Error bars denote bootstrap-estimated 95% confidence intervals. The dotted vertical lines denote the boundaries between early and late lists. **B.** In this panel, the fingerprints for early (circle) and late (star) lists are averaged across lists and participants before projecting the fingerprints into a (new) 2D space. See Figure S8 for analogous plots for the random conditions.

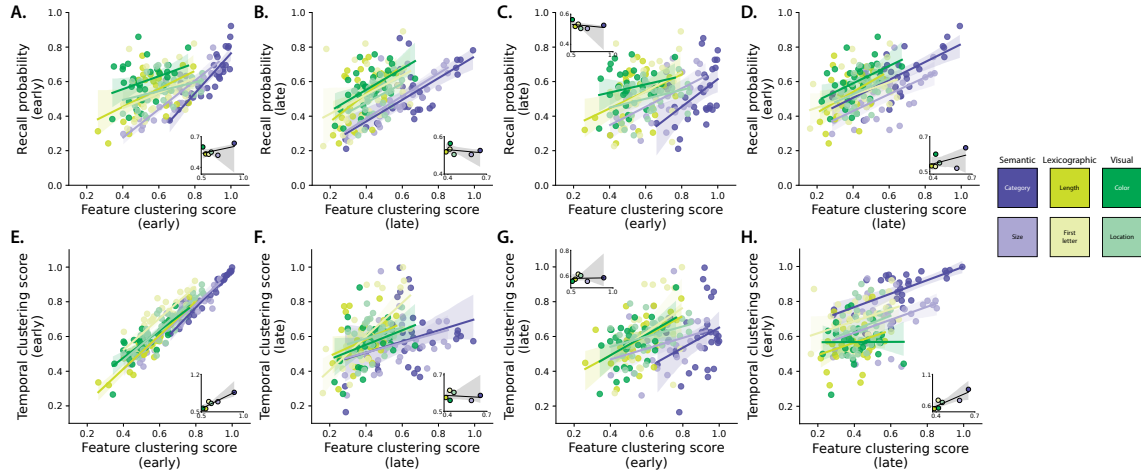
the order manipulations for those lists (see the locations of the circles in Fig. 6B). There also seemed to be some consistency across different features within a broader type. For example, both semantic feature conditions (category and size; purple markers) diverge in a similar direction from the group; 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 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 con-  
 ditions and participants (early:  $r(179) = 0.537, p < 0.001$ ; late:  $r(179) = 0.492, p < 0.001, CI = [0.352, 0.606]$ ;  
 late:  $r(179) = 0.492, p < 0.001$ ;  $r(179) = 0.403, p < 0.001, CI = [0.271, 0.517]$ ) as well as for  
 each condition individually for early ( $r_s \geq 0.386 \geq 0.331$ , all  $p_s \leq 0.035 \leq 0.069$ ) and late  
 ( $r_s \geq 0.462 \geq 0.404$ , all  $p_s \leq 0.010 \leq 0.027$ ) 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:  $r(4) = 0.526, p = 0.284$ ; late:  $r(4) = 0.511, p = 0.300, CI = [-0.999, 0.996]$ ; late:  $r(4) = -0.257, p = 0.623$ ;  
 see insets of Fig. 7A and B). We observed carryover associations between feature clustering  
 and recall performance (Fig. 7C, D). Participants who showed stronger feature clustering  
 on early lists in the non-visual conditions tended to recall more items on late lists (across  
 conditions:  $r(179) = 0.492, p < 0.001$ ; all  $r(179) = 0.230, p = 0.002, CI = [0.072, 0.372]$ ; all  
 non-visual conditions individually:  $r_s \geq 0.462 \geq 0.405$ , all  $p_s \leq 0.010 \leq 0.027$ ; color:  $r(29) = 0.212, p = 0.251, CI =$   
 location:  $r(28) = 0.320, p = 0.085, CI = [0.011, 0.584]$ ). Participants who recalled more  
 items on early lists also tended to show stronger feature clustering on late lists (across con-  
 ditions:  $r(179) = 0.280, p < 0.001$ ;  $r(179) = 0.464, p < 0.001, CI = [0.321, 0.582]$ ; individual  
 conditions: all non-visual conditions:  $r_s \geq 0.445 \geq 0.377$ , all  $p_s \leq 0.014$ ; color:  $r(29) = 0.298, p = 0.103$ ;  
 location:  $r(28) = 0.354, p = 0.055 \leq 0.040$ ). Neither of these effects showed condition-level  
 trends (early feature clustering versus late recall probability:  $r(4) = -0.299, p = 0.565$ ; early recall probability versus late feature clustering:  $r(4) = 0.400, p = 0.432$ ;  
 We also looked for associations between feature clustering and temporal clustering.  
 Across every order manipulation condition, participants who exhibited stronger fea-  
 ture clustering also exhibited stronger temporal clustering. For early lists (Fig. 7E), this  
 trend held overall ( $r(179) = 0.924, p < 0.001$ ;  $r(179) = 0.916, p < 0.001, CI = [0.893, 0.936]$ ),  
 for each condition individually (all  $r_s \geq 0.822$ , all  $p_s < 0.001$ ), and across conditions

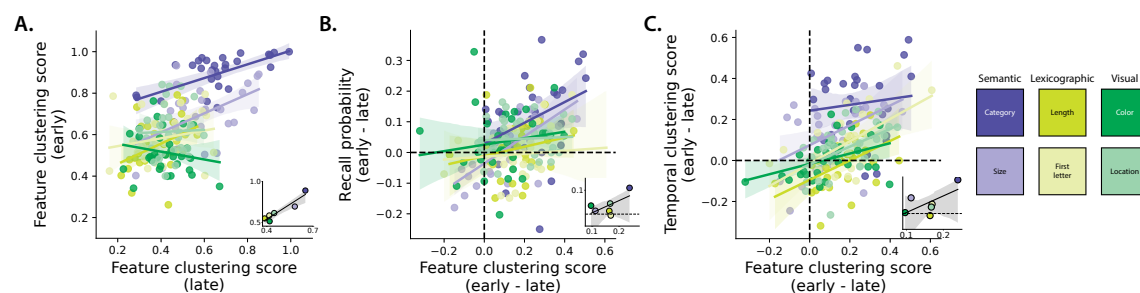
967 ( $r(4) = -0.964, p = 0.002$ ;  $r(4) = 0.964, p = 0.002$ ). For late lists (Fig. 7F), the results were more  
 968 variable (overall:  $r(179) = -0.348, p < 0.001$ ;  $r(179) = 0.348, p < 0.001$ ; all non-visual condi-  
 969 tions:  $r_s \geq 0.382$ , all  $p_s \leq 0.037$ ; color:  $r(29) = -0.453, p = 0.011$ ;  $r(29) = 0.453, p = 0.011$ ; loca-  
 970 tion:  $r(28) = -0.190, p = 0.314$ ;  $r(28) = 0.190, p = 0.314$ ; across-conditions:  $r(4) = -0.036, p = 0.945$ ;  $r(4) = -0.036, p = 0.945$ ).  
 971 While less robust than the carryover associations between feature clustering and re-  
 972 call performance, we also observed some carryover associations between feature clus-  
 973 tering and temporal clustering (Fig. 7G, H). Participants who showed stronger feature  
 974 clustering on early lists ~~trended towards showing~~ showed stronger temporal clustering  
 975 on later lists (overall:  $r(179) = -0.301, p < 0.001$ ;  $r(179) = 0.464, p < 0.001, CI = [0.321, 0.582]$ ;  
 976 for individual conditions: all  $r_s \geq -0.297 \geq 0.377$ , all  $p_s \leq 0.111 \leq 0.040$ ; across conditions:  
 977  $r(4) = 0.107, p = 0.840$ ;  $r(4) = 0.451, p = 0.369, CI = [-0.986, 0.998]$ ). And participants who  
 978 showed stronger temporal clustering on early lists ~~trended towards showing~~ stronger fea-  
 979 ture clustering on later lists (overall:  $r(179) = -0.579, p < 0.001$ ;  $r(179) = 0.266, p < 0.001, CI = [0.129, 0.396]$ ;  
 980 ~~for individual conditions: all non-visual conditions:  $r_s \geq 0.323$ , all  $p_s \leq 0.082$ ; visual~~  
 981 ~~conditions:  $r_s \geq 0.089 \geq 0.298$ , all  $p_s \leq 0.632 \leq 0.110$~~ ; across conditions:  $r(4) = -0.916, p = 0.010$ ;  $r(4) = 0.064, p = 0.945$ ).  
 982 Taken together, the results displayed in Figure 7 show that participants who were more  
 983 sensitive to the order manipulations (i.e., participants who showed stronger feature clus-  
 984 tering for their condition's feature on early lists) remembered more words and showed  
 985 stronger temporal clustering. These associations also appeared to carry over across lists,  
 986 even when the items on later lists were presented in a random order.

987 If participants show different sensitivities to order manipulations, how do their behav-  
 988 iors carry over to later lists? We found that participants who showed strong feature cluster-  
 989 ing on early lists often tended to show strong feature clustering on late lists (Fig. 8A; overall  
 990 across participants and conditions:  $r(179) = -0.592, p < 0.001$ ; ~~non-visual feature conditions:~~  
 991 ~~all  $r_s \geq 0.350$ , all  $p_s \leq 0.058$ ; color:  $r(29) = -0.071, p = 0.704$~~ ;  $r(179) = 0.591, p < 0.001, CI = [0.472, 0.682]$ ;  
~~visual conditions:  $r_s \geq 0.089 \geq 0.298$ , all  $p_s \leq 0.632 \leq 0.110$~~ ; across conditions:  $r(4) = -0.916, p = 0.010$ ;  $r(4) = 0.064, p = 0.945$ ).



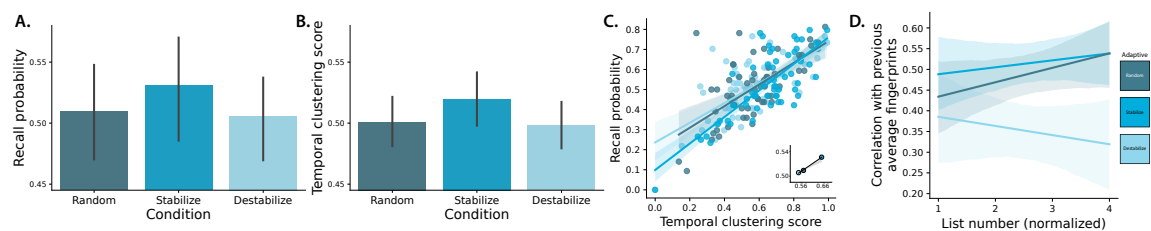
**Figure 7: Interactions between feature clustering, recall probability, and contiguity.** A. Recall probability versus feature clustering scores for order manipulation (early) lists. B. Recall probability versus feature clustering for randomly ordered (late) lists. C. Recall probability on late lists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

992 category:  $r(28) = 0.590, p < 0.001, CI = [0.354, 0.756]$ ; size:  $r(28) = 0.488, p = 0.006, CI = [0.134, 0.732]$ ;  
 993 length:  $r(28) = 0.384, p = 0.036, CI = [0.040, 0.681]$ ; first letter:  $r(28) = 0.202, p = 0.284, CI = [-0.273, 0.620]$ ;  
 994 color:  $r(29) = -0.183, p = 0.325, CI = [-0.562, 0.258]$ ; location:  ~~$r(28) = -0.032, p = -0.868$~~  $r(28) = 0.031, p = 0.870$ ;  
 995 across conditions:  ~~$r(4) = -0.934, p = -0.006$~~  $r(4) = 0.942, p = 0.005, CI = [0.442, 1.000]$ ). Al-  
 996 though participants tended to show weaker feature clustering on late lists (~~Fig. 6~~Figs. 4, 5)  
 997 on *average*, the associations between early and late lists for individual participants suggests  
 998 that some influence of early order manipulations may linger on late lists. We found that  
 999 participants who exhibited larger carryover in feature clustering (i.e., continued to show  
 1000 strong feature clustering on late lists) for the semantic order manipulations (but not other  
 1001 manipulations) also tended to show a ~~larger improvement in recall~~ smaller decrease in  
 1002 recall on early versus late lists (Fig. 8B; overall:  ~~$r(179) = -0.378, p < 0.001$~~  $r(179) = 0.307, p < 0.001, CI = [0.148, 0$   
 1003 category:  ~~$r(28) = -0.419, p = -0.021$~~  $r(28) = 0.350, p = 0.058, CI = [0.050, 0.642]$ ; size:  ~~$r(28) = -0.737, p < 0.001$ ;~~  
 1004 ~~non-semantic conditions: all  $r$ s  $\leq -0.252$ , all  $p$ s  $\geq 0.179$ ;~~ $r(28) = 0.708, p < 0.001, CI = [0.472, 0.862]$ ;  
 1005 length:  $r(28) = 0.205, p = 0.276, CI = [-0.109, 0.492]$ ; first letter:  $r(28) = 0.081, p = 0.672, CI = [-0.433, 0.597]$ ;  
 1006 color:  $r(29) = 0.155, p = 0.406, CI = [-0.174, 0.541]$ ; location:  $r(28) = 0.052, p = 0.787, CI = [-0.307, 0.360]$ ;  
 1007 across conditions:  ~~$r(4) = -0.773, p = -0.072$~~  on late lists, relative to early lists $r(4) = 0.635, p = 0.176, CI = [-0.924$   
 1008 Participants who exhibited larger carryover in feature clustering also tended to show  
 1009 stronger temporal clustering on late lists (relative to early lists) for all but the category con-  
 1010 dition (Fig. 8C; overall:  ~~$r(179) = -0.434, p < 0.001$~~  $r(179) = 0.426, p < 0.001, CI = [0.285, 0.544]$ ;  
 1011 category:  ~~$r(28) = -0.229, p = -0.223$~~  $r(28) = 0.110, p = 0.564, CI = [-0.284, 0.442]$ ; all non-category  
 1012 conditions: all  $r$ s  ~~$\geq -0.448$~~   $\geq 0.406$ , all  $p$ s  ~~$\leq -0.012$~~   $\leq 0.023$ ; across conditions:  ~~$r(4) = -0.598, p = -0.210$~~  $r(4) = 0.649, p =$   
 1013 We suggest two potential interpretations of these findings. First, it is possible that  
 1014 some participants are more “malleable” or “adaptable” with respect to how they organize  
 1015 incoming information. When presented with list of items sorted along *any* feature dimen-  
 1016 sion, they will simply adopt that feature as a dominant dimension for organizing those



**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.

1017 items and subsequent (randomly ordered) items. This flexibility in memory organization  
 1018 might afford such participants a memory advantage, explaining their strong recall perfor-  
 1019 mance. An alternative interpretation is that each participant comes into our study with a  
 1020 “preferred” way of organizing incoming information. If they happen to be assigned to an  
 1021 order manipulation condition that matches their preferences, then they will appear to be  
 1022 “sensitive” to the order manipulation and also exhibit a high degree of carryover in feature  
 1023 clustering from early to late lists. These participants might demonstrate strong recall per-  
 1024 formance not because of their inherently superior memory abilities, but rather because the  
 1025 specific condition they were assigned to happened to be especially easy for them, given  
 1026 their pre-experimental tendencies. To help distinguish between these interpretations, we  
 1027 designed an *adaptive* experimental condition (see *Adaptive condition*). The primary ma-  
 1028 nipulation in the adaptive condition is that participants each experience three key types  
 1029 of lists. On *random* lists, words are ordered randomly (as in the ~~feature-rich~~ feature-rich  
 1030 condition). On *stabilize* lists, the presentation order is adjusted to be maximally similar to



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

the current estimate of the participant's memory fingerprint (see *Online "fingerprint" analysis*). Third, on *destabilize* lists, the presentation order is adjusted to be *minimally* similar to the current estimate of the participant's memory fingerprint (see *Ordering "stabilize" and "destabilize" lists by an estimated fingerprint*). The orders in which participants experienced each type of list were counterbalanced across participants to help reduce the influence of potential list-order effects. Because the presentation orders on stabilize and destabilize lists are adjusted to best match each participant's (~~potentially-unique~~) ~~personal~~ 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 remember-

ing words on stabilize lists relative to words on both random ( $t(59) = 1.740, p = 0.087, d = 0.114$ ) and destabilize ( $t(59) = 1.714, p = 0.092, d = 0.114, CI = [-0.351, 4.108]$ ) lists (Fig. 9A). Participants showed no reliable differences in their memory performance on destabilize versus random lists ( $t(59) = -0.249, p = 0.804, d = -0.017, CI = [-2.327, 1.959]$ ). Participants also exhibited stronger temporal clustering on stabilize lists, relative to random ( $t(59) = 3.554, p = 0.001, d = 0.306, CI = [1.635, 5.460]$ ) and destabilize ( $t(59) = 4.045, p < 0.001, d = 0.374, CI = [1.964, 6.968]$ ) lists (Fig. 9B). We found no reliable differences in temporal clustering for items on random versus destabilize lists ( $t(59) = -0.781, p = 0.438, d = -0.081, CI = [-3.165, 1.127]$ ).

As in the other experimental manipulations, participants in the adaptive condition exhibited substantial variability with respect to their overall memory performance and their clustering tendencies (Fig. 9C). We found that individual participants who exhibited strong temporal clustering scores also tended to recall more items. This held across subjects, aggregating across all list types ( $r(178) = 0.721, p < 0.001, CI = [0.590, 0.789]$ ), and for each list type individually (all  $rs \geq 0.683 \geq 0.651$ , all  $ps \leq 0.001 < 0.001$ ). Taken together, the results from the adaptive condition suggest that each participant comes into the experiment with their own unique memory organization tendencies, as characterized by their memory fingerprint. When participants study lists whose items come pre-sorted according to their unique preferences, they tend to remember more and show stronger temporal clustering.

We note that the multivariate aspect of the adaptive condition (i.e., sorting lists simultaneously along multiple feature dimensions) provides an important contrast with the other order manipulation conditions, where we sort lists along only a single feature dimension in each condition. We found that participants “naturally” clustered their recalls along multiple feature dimensions, even when the lists they studied were not sorted along



	<i>t</i> -value	df	Cohen's <i>d</i>	<i>p</i> -value (raw)	<i>p</i> -value (corrected)	95% CI (lower bound)	95% CI (upper bound)
Rank 1	12.751	66	0.162	< 0.001	< 0.001	8.702	20.013
Rank 2	8.196	66	0.162	< 0.001	< 0.001	4.794	12.978
Rank 3	3.243	66	0.162	0.002	0.002	1.028	7.051
Rank 4	-3.112	66	0.162	0.003	0.003	-5.282	-1.920
Rank 5	-7.154	66	0.162	< 0.001	< 0.001	-12.649	-5.568
Rank 6	-12.608	66	0.162	< 0.001	< 0.001	-22.114	-9.347
Rank 7	-18.397	66	0.162	< 0.001	< 0.001	-27.238	-14.073

**Table 23: Ranked clustering scores versus “chance” for participants in the feature-rich condition.** For each participant, we sorted their clustering scores in descending order (for each of the six feature dimensions, along with a seventh dimension to capture temporal clustering). The *t*-tests reported in the table (for the clustering scores at each “rank”) were carried out across-participants, and reflect data aggregated across all lists from each participant. Abbreviations used in this table are defined in Table S1.

those dimensions (as in the feature-rich condition). A caveat is that the *specific* feature dimensions participants tended to cluster along varied across participants. One way to quantify the multidimensional nature of participants’ clustering tendencies is to sort each participant’s clustering scores (for each of the six feature dimensions, along with a seventh dimension to capture temporal clustering). We can then ask whether the distribution of clustering scores at each “rank” within the sorted set of scores for each participant has a mean that is reliably different from a chance value of 0.5. We carried out these tests for each set of ranked scores, and found that participants in the feature-rich condition reliably cluster their recalls along at least three dimensions, including temporal clustering (which was often ranked highest; Tab. 23). That the clustering scores ranked in the top three dimensions were reliably greater than chance suggests that participants organize their memories along at least three feature dimensions, even when the words are studied in a random order.

## Discussion

We asked participants to study and freely recall word lists. The words on each list (and the total set of lists) were held constant across participants. For each word, we considered

1086 (and manipulated) two semantic features (category and size) that reflected aspects of the  
1087 *meanings* of the words, along with two lexicographic features (word length and first letter),  
1088 which reflected characteristics of the words' *letters*. These semantic and lexicographic  
1089 features are intrinsic to each word. We also considered and manipulated two additional  
1090 visual features (color and location) that affected the *appearance* of each studied item, but  
1091 could be varied independently of the words' identities. Across different experimental  
1092 conditions, we manipulated how the visual features varied across words (within each  
1093 list), along with the orders of each list's words. Although the participants' task (verbally  
1094 recalling as many words as possible, in any order, within one minute) remained constant  
1095 across all of these conditions, and although the set of words they studied from each list  
1096 remained constant, our manipulations substantially affected participants' memories. The  
1097 impact of some of the manipulations also affected how participants remembered *future*  
1098 lists that were sorted randomly.

#### 1099 **Recap: visual feature manipulations**

1100 We found that participants in our ~~feature-rich~~feature-rich condition (where we varied  
1101 words' appearances) recalled similar proportions of words to participants in a reduced  
1102 condition (where appearance was held constant across words). However, varying the  
1103 words' appearances led participants to exhibit much more temporal and feature-based  
1104 clustering. This suggests that even seemingly irrelevant elements of our experiences can  
1105 affect how we remember them.

1106 When we held the within-list variability in participants' visual experiences fixed across  
1107 lists (in the ~~feature-rich~~feature-rich and reduced conditions), they remembered more words  
1108 from early lists than from late lists. For ~~feature-rich~~feature-rich lists, they also showed  
1109 stronger clustering for early versus late lists. However, when we *varied* participants' visual

1110 experiences across lists (in the “reduced (early)” and “reduced (late)” conditions), these  
1111 early versus late accuracy and clustering differences disappeared. Abruptly changing  
1112 how incidental visual features varied across words seemed to act as a sort of “event  
1113 boundary” that partially reset how participants processed and remembered post-boundary  
1114 lists. Within-list clustering also increased in these manipulations, suggesting that the  
1115 “within-event” words were being more tightly associated with each other.

1116 When we held the visual features constant during early lists, but then varied words’  
1117 appearances in later lists (i.e., the reduced (early) condition), participants’ overall memory  
1118 performance improved. However, this impact was directional: when we *removed* visual  
1119 features from words in late lists that had been present in early lists (i.e., the reduced (late)  
1120 condition), we saw no memory improvement.

## 1121 **Recap: order manipulations**

1122 When we (stochastically) sorted early lists along different feature dimensions, we found  
1123 several impacts on participants’ memories. Sorting early lists semantically (by word cat-  
1124 egory) enhanced participants’ memories for those lists, but the effects on performance of  
1125 sorting along other feature dimensions were inconclusive. However, each order manipu-  
1126 lation substantially affected how participants *organized* their memories of words from the  
1127 ordered lists. When we sorted lists semantically, participants displayed stronger semantic  
1128 clustering; when we sorted lists lexicographically, they displayed stronger lexicographic  
1129 clustering; and when we sorted lists visually, they displayed stronger visual clustering.  
1130 Clustering along the unmanipulated feature dimensions in each of these cases was un-  
1131 changed.

1132 The order manipulations we examined also appeared to induce, in some cases, a  
1133 tendency to “clump” similar words within a list. This was most apparent on semantically

1134 ordered lists, where the probability of initiating recall with a given word seemed to follow  
1135 groupings defined by feature change points.

1136 We also examined the impact of early list order manipulations on memory for late  
1137 lists. At the group level, we found little evidence for lingering “carryover” effects of these  
1138 manipulations: participants in the order manipulation conditions showed similar mem-  
1139 ory performance and clustering on late lists to participants in the corresponding control  
1140 (~~feature-rich~~feature-rich) condition. At the level of individual participants, however, we  
1141 found several meaningful patterns.

1142 Participants who showed stronger feature clustering on early (order-manipulated) lists  
1143 tended to better remember late (randomly ordered) lists. Participants who remembered  
1144 early lists better also tended to show stronger feature clustering (along their condition’s  
1145 feature dimension) on late lists (even though the words on those late lists were presented  
1146 in a random order). We also observed some (weaker) carryover effects of temporal cluster-  
1147 ing. Participants who showed stronger feature clustering (along their condition’s feature  
1148 dimension) on early lists tended to show stronger temporal clustering on late lists. And  
1149 participants who showed stronger temporal clustering on early lists also tended to show  
1150 stronger feature clustering on late lists. Essentially, these order manipulations appeared to  
1151 affect each participant differently. Some participants were sensitive to our manipulations,  
1152 and those participants’ memory performance was impacted more strongly, both for the  
1153 ordered lists and for future (random) lists. Other participants appeared relatively insen-  
1154 sitive to our manipulations, and those participants showed little carryover effects on late  
1155 lists.

1156 These results at the individual participant level suggested to us that either (a) some  
1157 participants were more sensitive to *any* order manipulation, or (b) some participants might  
1158 be more (or less) sensitive to manipulations along *particular* (e.g., preferred) feature dimen-

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

### Memory consequences of feature variability

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

1183 may be more important than the absolute complexity or numbers of meaningful features.  
1184 Some prior studies have suggested that people can “cue” their memories using different  
1185 “strategies” or “pathways” for searching for the target information. For example, modern  
1186 accounts of free recall often posit that memory search typically begins by matching  
1187 the current state of mental context with the contexts associated with other items in  
1188 memory (Kahana, 2020). Since context is the defining hallmark of episodic memory (Tulving, 1983)  
1189 , context-based search can be described as an “episodic” pathway to recall. When episodic  
1190 cueing fails to elicit a match, participants may then search for items that are similar to  
1191 the current mental context or mental state along other dimensions, such as semantic  
1192 similarity (Davachi et al., 2003; Socher et al., 2009). These multiple pathways accounts of  
1193 memory search also provide a potential explanation of why participants might have an  
1194 easier time remembering richer stimuli (or experiences): richer stimuli and experiences  
1195 might have more features that could be used to cue memory search. Our work suggests  
1196 that there may be some additional factors at play with respect to the *dynamics* of these  
1197 processes. In particular, we only observed memory benefits for “richer” stimuli when they  
1198 were encountered after more “impoverished” stimuli (in the reduced (early) condition).  
1199 This suggests that the pathways available to recall a given item may also depend on recent  
1200 prior experiences.

1201 We did *not* find any evidence that changing words’ appearances *harmed* memory  
1202 performance, e.g., by distracting them with irrelevant information (Lange, 2005; Marsh et al., 2012, 2015; Reini  
1203 . Nor did we find any evidence that *changes* in the presence of potentially “distracting”  
1204 features adversely affected memory. For example, when we increased or decreased the  
1205 variability in words’ appearances on late versus early lists (as in the reduced (early) and  
1206 reduced (late) conditions), we found no evidence that this harmed participants’ memories.  
1207 One potential interpretation under the “multiple pathways to recall” framework is that

1208 the availability of multiple pathways to recall do not appear to specifically interfere with  
1209 each other.

## 1210 **Context effects on memory performance and organization**

1211 In real-world ~~experience~~experiences, each moment's unique blend of contextual features  
1212 (where we are, who we are with, what else we are thinking of at the time, what else we  
1213 experience nearby in time, etc.) plays an important role in how we interpret, experience,  
1214 and remember that moment, and how we relate it to our other experiences (e.g., for review  
1215 see Manning, 2020). What are the analogues of real-world contexts in laboratory tasks like  
1216 the free recall paradigm employed in our study? In general, modern formal accounts of  
1217 free recall (Kahana, 2020) describe context as comprising a mix of (a) features pertaining to  
1218 or associated with each item and (b) other items and thoughts experienced nearby in time,  
1219 e.g., that might still be "lingering" in the participant's thoughts at the time they study  
1220 the item. Item features can include semantic properties (i.e., features related to the item's  
1221 meaning), lexicographic properties (i.e., features related to the item's letters), sensory  
1222 properties (i.e., feature related to the item's appearance, sound, smell, etc.), emotional  
1223 properties (i.e., features related to how meaningful the item is, whether the item evokes  
1224 positive or negative feelings, etc.), utility-related properties (e.g., features that describe  
1225 how an item might be used or incorporated into a particular task or situation), and more.  
1226 Essentially any aspect of the participant's experience that can be characterized, measured,  
1227 or otherwise described can be considered to influence the participant's mental context at  
1228 the moment they experience that item. Temporally proximal features include aspects of  
1229 the participant's internal or external experience that are *not* specifically occurring at the  
1230 moment they encounter an item, but that nonetheless influence how they process the item.  
1231 Thoughts related to percepts, goals, expectations, other experiences, and so on that might

1232 have been cued (directly or indirectly) by the participant’s recent experiences prior to the  
1233 current moment all fall into this category. Internally driven mental states, such as thinking  
1234 about an experience unrelated to the experiment, also fall into this category.

1235 Contextual features need not be intentionally or consciously perceived by the partic-  
1236 ipant to affect memory, nor do they need to be relevant to the task instructions or the  
1237 participant’s goals. Incidental factors such as font color (Jones and Pyc, 2014), back-  
1238 ground color (Isarida and Isarida, 2007), inter-stimulus images (Chiu et al., 2021; Gersh-  
1239 man et al., 2013; Manning et al., 2016), background sounds (~~Sahakyan and Smith, 2014; ?~~)  
1240 (Sahakyan and Smith, 2014; ?), secondary tasks (Masicampto and Sahakyan, 2014; Ober-  
1241 auer and Lewandowsky, 2008; Polyn et al., 2009), and more can all impact how participants  
1242 remember, and organize in memory, lists of studied items.

1243 Consistent with this prior work, we found that participants were sensitive to task-  
1244 irrelevant visual features. We also found that changing the dynamics of those task-  
1245 irrelevant visual features (in the reduced (early) and reduced (late) conditions) *also* affected  
1246 participants’ memories. This suggests that it is not only the contextual features themselves  
1247 that affect memory, but also the *dynamics* of context—i.e., how the contextual features  
1248 associated with each item change over time.

## 1249 **Priming effects on memory performance and organization**

1250 When our ongoing experiences are ambiguous, we can draw on our past experiences,  
1251 expectations, and other real, perceived, or inferred cues to help resolve these ambiguities.  
1252 We may also be overtly or covertly “primed” to influence how we are likely to resolve  
1253 ambiguities. For example, before listening to a story with several equally plausible inter-  
1254 pretations, providing participants with “background” information beforehand can lead  
1255 them towards one interpretation versus another (Yeshurun et al., 2017). More broadly, our



conscious and unconscious biases and preferences can influence not only how we interpret high-level ambiguities, but even how we process low-level sensory information (Katabi et al., 2023).

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

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

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

A large number of prior studies have compared participants’ memories for categorized word lists that are presented in blocked versus random orders. In “blocked” lists, all of the words from a given semantic category (e.g., animals) are presented consecutively,

1280 whereas in “random” lists, the words from different categories are intermixed. Most of  
1281 these studies report that participants tend to better remember blocked (versus random)  
1282 lists (Bower et al., 1969; Cofer et al., 1966; D’Agostino, 1969; Dallett, 1964; Kintsch, 1970; Luek et al., 1971; Pu  
1283 . Other studies suggest that these order effects may also be modulated by factors like list  
1284 length and the numbers of exemplars in each category (e.g., Borges and Mangler, 1972).

1285 Although we did not directly manipulate “blocking” in our order manipulation conditions,  
1286 our sorting procedures in those conditions (see *Constructing feature-sorted lists*) have  
1287 indirect effects on the lists’ blockiness. For example, lists that are stochastically sorted by  
1288 semantic category will tend to contain runs of several same-category words in succession.  
1289 Consistent with the above work on blocked versus random categorized lists, we found  
1290 that participants tended to better remember lists that were sorted semantically (Fig. 5B).  
1291 However, this memory improvement did not appear to extend to the other order manipulation  
1292 conditions we considered (e.g., to lexicographically or visually sorted lists). One possibility  
1293 is that the memory benefits of blocked versus random lists are specific to semantic  
1294 categories, and do not generalize to other feature dimensions. Another possibility is that  
1295 the memory benefits are due to the presence of infrequent “jumps” between successive  
1296 items (e.g., from different categories). Because the features we manipulated in the  
1297 lexicographic and visual conditions were less categorical than the semantic features,  
1298 feature values across words in those conditions tended to vary more gradually. Relatively  
1299 stable features that are punctuated by infrequent large changes (e.g., as words transition  
1300 from a same-category sequence to a new category) may also relate to perceived “event  
1301 boundaries,” which can have important consequences for memory (DuBrow and Davachi, 2013, 2016; DuBrow  
1302 ~

## 1303 **Expectation, event boundaries, and situation models**

1304 Our findings that participants' current and future memory behaviors are sensitive to  
1305 manipulations in which features change over time, and how features change across items  
1306 and lists, suggest parallels with studies on how we form expectations and predictions,  
1307 segment our continuous experiences into discrete events, and make sense of different  
1308 scenarios and situations. Each of these real-world cognitive phenomena entail identifying  
1309 statistical regularities in our experiences, and exploiting those regularities to gain insight,  
1310 form inferences, organize or interpret memories, and so on. Our past experiences enable  
1311 us to predict what is likely to happen in the future, given what happened "next" in our  
1312 previous experiences that were similar to now (Barron et al., 2020; Brigard, 2012; Chow  
1313 et al., 2016; Eichenbaum and Fortin, 2009; Gluck et al., 2002; Goldstein et al., 2021; Griffiths  
1314 and Steyvers, 2003; Jones and Pashler, 2007; Kim et al., 2014; Manning, 2020; Tamir and  
1315 Thornton, 2018; Xu et al., 2023).

1316 When our expectations are violated, such as when our observations disagree with our  
1317 predictions, we may perceive the "rules" or "situation" to have changed. *Event boundaries*  
1318 denote abrupt changes in the state of our experience, for example, when we transition  
1319 from one situation to another (Radvansky and Zacks, 2017; Zwaan and Radvansky, 1998).  
1320 Crossing an event boundary can impair our memory for pre-boundary information and en-  
1321 hance our memory for post-boundary information (DuBrow and Davachi, 2013; Manning  
1322 et al., 2016; Radvansky and Copeland, 2006; Sahakyan and Kelley, 2002). Event bound-  
1323 aries are also tightly associated with the notion of *situation models* and *schemas*—mental  
1324 frameworks for organizing our understanding about the rules of how we and others are  
1325 likely to behave, how events are likely to unfold over time, how different elements are  
1326 likely to interact, and so on. For example, a situation model pertaining to a particular  
1327 restaurant might set our expectations about what we are likely to experience when we

1328 visit that restaurant (e.g., what the building will look like, how it will smell when we enter,  
1329 how crowded the restaurant is likely to be, the sounds we are likely to hear, etc.). Similarly,  
1330 as mentioned in the *Introduction*, we might learn a schema describing how events are likely  
1331 to unfold *across* any sit-down restaurant—e.g., open the door, wait to be seated, receive a  
1332 menu, decide what to order, place the order, and so on. Situation models and schemas can  
1333 help us to generalize across our experiences, and to generate expectations about how new  
1334 experiences are likely to unfold. When those expectations are violated, we can perceive  
1335 ourselves to have crossed into a new situation.

1336 In our study, we found that abruptly changing the “rules” about how the visual  
1337 appearances of words are determined, or about the orders in which words are presented,  
1338 can lead participants to behave similarly to what one might expect upon crossing an event  
1339 boundary. Adding variability in font color and presentation location for words on late  
1340 lists, after those visual features had been held constant on early lists, led participants to  
1341 remember more words on those later lists. One potential explanation is that participants  
1342 perceive an “event boundary” to have occurred when they encounter the first “late” list.  
1343 According to contextual change accounts of memory across event boundaries (e.g., Flores  
1344 et al., 2017; Gold et al., 2017; Pettijohn et al., 2016; Sahakyan and Kelley, 2002), this could  
1345 help to explain why participants in the reduced (early) condition exhibited better overall  
1346 memory performance. Specifically, their memory for late list items could benefit from less  
1347 interference from early list items, and the contextual features associated with late list items  
1348 (after the “event boundary”) might serve as more specific recall cues for those late items  
1349 (relative to if the boundary had not occurred).

1350 How do different types of clustering relate to each other, and to memory  
1351 performance?

1352 When the words on a studied list are presented in a random order, different types of  
1353 clustering in participants' recalls often tend to be negatively correlated. For example,  
1354 words that occur nearby on the list will not (on average) tend to be semantically related, and  
1355 vice versa. Therefore a participant who shows a strong tendency to temporally cluster their  
1356 recalls will tend to show weaker semantic clustering, and so on (Healey and Uitvlugt, 2019; Howard and Kahana  
1357 . Further, there is some evidence that temporal clustering is positively correlated with  
1358 memory performance, whereas semantic clustering is negatively correlated with memory  
1359 performance (Sederberg et al., 2010).

1360 The notion of "multiple pathways to recall" discussed above (see *Memory consequences*  
1361 *of feature variability*) suggests one potential explanation for these patterns. For example,  
1362 temporal clustering has been proposed to reflect reliance on contextual cues in an "episodic"  
1363 pathway to search memory, whereas semantic clustering reflects a relies on specific item  
1364 features. These two pathways may "compete" with each other during recall (Socher et al., 2009)  
1365 . Meanwhile, extra-list intrusion errors (i.e., false "recalls" of items that were never  
1366 encountered on the list) often tend to share semantic features with recently recalled  
1367 items (Zaromb et al., 2006) and also often lead the participant to stop recalling additional  
1368 items (Miller et al., 2012). Speculatively, over-reliance on semantic cues may lead to more  
1369 intrusion errors, which in turn may lead to fewer recalls overall.

1370 Our findings extend these prior results to consider lists that are *not* ordered randomly.  
1371 Because ordering the words on a list along a particular feature dimension removes the  
1372 "conflict" between temporal and feature clustering, the order manipulation conditions  
1373 in our study represent an "edge case" whereby different pathways to recall are not  
1374 necessarily in conflict with each other. For example, the same participants who exhibit  
1375 strong feature clustering *also* show strong temporal clustering on ordered lists (Fig. 7E).  
1376 This is presumably at least partly due to an inability to separate temporal and feature

1377 clustering on ordered lists (also see *Factoring out the effects of temporal clustering*). However,  
1378 features that change gradually with time (i.e., presentation position) could also serve  
1379 to strengthen the episodic (contextual) cues associated with each item. In other words,  
1380 participants might essentially combine multiple noisy measures of change to form a more  
1381 stable internal representation of temporal context.

## 1382 **Theoretical implications**

1383 Although most modern formal theories of episodic memory have been developed and  
1384 tested to explain memory for list-learning tasks (Kahana, 2020), a number of recent studies  
1385 suggest some substantial differences between memory for lists versus naturalistic stim-  
1386 uli (e.g., real-world experiences, narratives, films, etc.; Heusser et al., 2021; Lee et al., 2020;  
1387 Manning, 2021; Nastase et al., 2020). One reason is that naturalistic stimuli are often much  
1388 more engaging than the highly simplified list-learning tasks typically employed in the  
1389 psychological laboratory, perhaps leading participants to pay more attention, exert more  
1390 effort, and stay more consistently motivated to perform well (Nastase et al., 2020). Another  
1391 reason is that the temporal unfoldings of events and occurrences in naturalistic stimuli  
1392 tend to be much more meaningful than the temporal unfoldings of items on typical lists  
1393 used in laboratory memory tasks. Real-world events exhibit important associations at a  
1394 broad range of timescales. For example, an early detail in a detective story may prove to  
1395 be a clue to solving the mystery later on. Further, what happens in one moment typically  
1396 carries some predictive information about what came before or after (Xu et al., 2023). In  
1397 contrast, the lists used in laboratory memory tasks are most often ordered randomly, by  
1398 design, to *remove* meaningful temporal structure in the stimulus (Kahana, 2012).

1399 On one hand, naturalistic stimuli provide a potential means of understanding how our  
1400 memory systems function in the circumstances we most often encounter in our everyday

1401 lives. This implies that, to understand how memory works in the “real world,” we should  
1402 study memory for stimuli that reflect the relevant statistical structure of real-world expe-  
1403 riences. On the other hand, naturalistic stimuli can be difficult to precisely characterize or  
1404 model, making it difficult to distinguish whether specific behavioral trends follow from  
1405 fundamental workings of our memory systems, from some aspect of the stimulus, or from  
1406 idiosyncratic interactions or interference between participants’ memory systems and the  
1407 stimulus. This challenge implies that, to understand the fundamental nature of memory  
1408 in its “pure” form, we should study memory for highly simplified stimuli that can pro-  
1409 vide relatively unbiased (compared with real-world experiences) measures of the relevant  
1410 patterns and tendencies.

1411     The experiment we report in this paper was designed to help bridge some of this gap  
1412 between naturalistic tasks and more traditional list-learning tasks. We had people study  
1413 word lists similar to those used in classic memory studies, but we also systematically var-  
1414 ied the lists’ “richness” (by adding or removing visual features) and temporal structure  
1415 (through order manipulations that varied over time and across experimental conditions).  
1416 We found that participants’ memory behaviors were sensitive to these manipulations.  
1417 Some of the manipulations led to changes that were common across people (e.g., more  
1418 temporal clustering when words’ appearances were varied, enhanced memory for lists  
1419 following an “event boundary,” more feature clustering on order-manipulated lists, etc.).  
1420 Other manipulations led to changes that were idiosyncratic (especially carryover effects  
1421 from order manipulations; e.g., participants who remembered more words on early order-  
1422 manipulated lists tended to show stronger feature clustering for their condition’s feature  
1423 dimension on late randomly ordered lists, etc.). We also found that participants remem-  
1424 bered more words from lists that were sorted to align with their idiosyncratic clustering  
1425 preferences. Taken together, our results suggest that our memories are susceptible to ex-

1426 ternal influences (i.e., to the statistical structure of ongoing experiences), but the effects of  
1427 past experiences on future memory are largely idiosyncratic across people.

## 1428 **Potential applications**

1429 Every participant in our study encountered exactly the same words, split into exactly the  
1430 same lists. But participants' memory performance, the orders in which they recalled the  
1431 words, and the effects of early list manipulations on later lists all varied according to how  
1432 we presented the to-be-remembered words.

1433 Our findings raise a number of exciting questions. For example, how far might these  
1434 manipulations be extended? In other words, might there be more sophisticated or clever  
1435 feature or order manipulations that one could implement to have stronger impacts on  
1436 memory? Are there limits to how much impact (on memory performance and/or or-  
1437 ganization) these sorts of manipulations can have? Are those limits universal across  
1438 people, or are there individual differences (based on prior experiences, natural strate-  
1439 gies, neuroanatomy, etc.) that impose person-specific limits on the potential impact of  
1440 presentation-level manipulations on memory?

1441 Our findings indicate that the ways word lists are presented affects how people re-  
1442 member them. To the extent that word list memory reflects memory processes that are  
1443 relevant to real-world experiences, one could imagine potential real-world applications of  
1444 our findings. For example, we found that participants remembered more words when the  
1445 presentation order agreed with their memory fingerprints. If analogous fingerprints could  
1446 be estimated for classroom content, perhaps they could be utilized manually by teachers,  
1447 or even by automated content-presentation systems, to optimize how and what students  
1448 remember.



## 1449 **Concluding remarks**

1450 Our work raises deep questions about the fundamental nature of human learning. What  
1451 are the limits of our memory systems? How much does what we remember (and how we  
1452 remember) depend on how we learn or experience the to-be-remembered content? We  
1453 know that our expectations, strategies, situation models learned through prior experiences,  
1454 and more collectively shape how our experiences are remembered. But those aspects of  
1455 our memory are not fixed: when we are exposed to the same experience in a new way, it  
1456 can change how we remember that experience, and also how we remember, process, or  
1457 perceive *future* experiences.

## 1458 **Author contributions**

1459 Conceptualization: JRM and ACH. ~~Methodology: JRM and ACH. Software: JRM~~  
1460 ~~curation: JRM, PCF, CEF, and ACH. Analysis~~ACH. ~~Formal analysis:~~ JRM, PCF, and ACH.  
1461 ~~Data collection: Funding acquisition: JRM. Investigation:~~ ECW, PCF, MRL, AMF, BJB,  
1462 DR, ~~and CEF. Data curation and management: ECW~~ CEF, and ACH. ~~Methodology: JRM~~  
1463 ~~and ACH. Project administration: ECW and PCF. Resources: JRM. Software: JRM, PCF,~~  
1464 ~~MRL~~ CEF, and ACH. ~~Supervision: JRM and ACH. Validation: JRM, PCF, and ACH.~~ Writing  
1465 (original draft): JRM. Writing (review and editing): ECW, PCF, MRL, AMF, BJB, DR, CEF,  
1466 and ACH. ~~Supervision: JRM and ACH. Project administration: ECW and PCF. Funding~~  
1467 ~~acquisition: JRM.~~

## 1468 ~~Data and code availability~~ Author note

1469 All of the data analyzed in this manuscript, along with all of the code for carrying out the  
1470 analyses may be found at <https://github.com/ContextLab/FRFR-analyses>. Code for run-  
1471 ning the non-adaptive experimental conditions may be found at [73](https://github.com/Con-</a></p></div><div data-bbox=)

1472 textLab/efficient-learning-code. Code for running the adaptive experimental condition  
1473 may be found at <https://github.com/ContextLab/adaptiveFR>. We have also released an as-  
1474 sociated Python toolbox for analyzing free recall data, which may be found at [https://cdl-](https://cdl-quail.readthedocs.io/en/latest/)  
1475 [quail.readthedocs.io/en/latest/](https://cdl-quail.readthedocs.io/en/latest/). Note that this study was not preregistered. Some of the  
1476 ideas and data presented in this manuscript were also presented at the Annual Meeting  
1477 of the Society for Neuroscience (2017) and the Context and Episodic Memory Symposium  
1478 (2017).

## 1479 Acknowledgements

1480 We acknowledge useful discussions, assistance in setting up an earlier (unpublished)  
1481 version of this study, and assistance with some of the data collection efforts from Rachel  
1482 Chacko, Joseph Finkelstein, Sheherzad Mohyidin, Lucy Owen, Gal Perlman, Jake Rost,  
1483 Jessica Tin, Marisol Tracy, Peter Tran, and Kirsten Ziman. Our work was supported in part  
1484 by NSF CAREER Award Number 2145172 to JRM. The content is solely the responsibility  
1485 of the authors and does not necessarily represent the official views of our supporting  
1486 organizations. The funders had no role in study design, data collection and analysis,  
1487 decision to publish, or preparation of the manuscript.

## 1488 References

- 1489 Anderson, J. R. and Bower, G. H. (1972). Recognition and retrieval processes in free recall.  
1490 *Psychological Review*, 79(2):97–123.
- 1491 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its  
1492 control processes. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning*  
1493 *and Motivation*, volume 2, pages 89–105. Academic Press, New York, NY.

- 1494 Baddeley, A. D. (1968). Prior recall of newly learned items and the recency effect in free  
1495 recall. *Canadian Journal of Psychology*, 22:157–163.
- 1496 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event  
1497 schemas during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.
- 1498 Balota, D. A., Black, S. R., and Cheney, M. (1992). Automatic and attentional priming in  
1499 young and older adults: reevaluation of the two-process model. *Journal of Experimental*  
1500 *Psychology: Human Perception and Performance*, 18(2):485–502.
- 1501 Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a  
1502 predictive coding account. *Progress in Neurobiology*, 192:101821–101834.
- 1503 Bonin, P., Thiebaut, G., Bugaiska, A., and Méot, A. (2022). Mixed evidence for a richness-of-  
1504 encoding account of animacy effects in memory from the generation-of-ideas paradigm.  
1505 *Current Psychology*, 41:1653–1662.
- 1506 Borges, M. A. and Mangler, G. (1972). Effect of within-category spacing on free recall.  
1507 *Journal of Experimental Psychology*, 92(2):207–214.
- 1508 Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged  
1509 associates. *Journal of General Psychology*, 49:229–240.
- 1510 Bousfield, W. A., Sedgewick, C. H., and Cohen, B. H. (1954). Certain temporal character-  
1511 istics of the recall of verbal associates. *American Journal of Psychology*, 67:111–118.
- 1512 Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive*  
1513 *Psychology*, 11(2):177–220.
- 1514 Bower, G. H., Lesgold, A. M., and Tieman, D. (1969). Grouping operations in free recall.  
1515 *Journal of Verbal Learning and Verbal Behavior*, 8(4):481–493.

- 1516 Brigard, F. D. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*,  
1517 3(420):1–3.
- 1518 Chiu, Y.-C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., and Sahakyan, L. (2021). Sepa-  
1519 ration of item and context in item-method directed forgetting. *NeuroImage*, 235:117983.
- 1520 Chow, W.-Y., Momma, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory  
1521 retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.
- 1522 Clayton, K. and Chattin, D. (1989). Spatial and semantic priming effects in tests of spa-  
1523 tial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,  
1524 15(3):495–506.
- 1525 Clewett, D., DuBrow, S., and Davachi, L. (2019). Transcending time in the brain: how  
1526 event memories are constructed from experience. *Hippocampus*, 29(3):162–183.
- 1527 Cofer, C. N., Bruce, D. R., and Reicher, G. M. (1966). Clustering in free recall as a function  
1528 of certain methodological variations. *Journal of Experimental Psychology: General*, 71:858–  
1529 866.
- 1530 D’Agostino, P. R. (1969). The blocked-random effect in recall and recognition. *Journal of*  
1531 *Verbal Learning and Verbal Behavior*, 8:815–820.
- 1532 Dallett, K. M. (1964). Number of categories and category information in free recall. *Journal*  
1533 *of Experimental Psychology*, 68:1–12.
- 1534 Darley, C. F. and Murdock, B. B. (1971). Effects of prior free recall testing on final recall  
1535 and recognition. *Journal of Experimental Psychology: General*, 91:66–73.
- 1536 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct

- medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences, USA*, 100(4):2157–2162.
- Donnelly, R. E. (1988). Priming effects in successive episodic tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:256–265.
- Drewnowski, A. and Murdock, B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory*, 6:319–332.
- DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- DuBrow, S. and Davachi, L. (2016). Temporal binding within and across events. *Neurobiology of Learning and Memory*, 134:107–114.
- DuBrow, S., Rouhani, N., Niv, Y., and Norman, K. A. (2017). Does mental context drift or shift? *Current Opinion in Behavioral Sciences*, 17:141–146.
- Eichenbaum, H. and Fortin, N. J. (2009). The neurobiology of memory based predictions. *Philosophical Transactions of the Royal Society of London Series B*, 364(1521):1183–1191.
- Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, 62:145–154.
- Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science*, 22(2):243–252.
- Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall: a challenge for unitary models of recency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36:324–347.

- 1559 Farrell, S. (2014). Correcting the correction of conditional recency slopes. *Psychonomic*  
1560 *Bulletin and Review*, 21:1174–1179.
- 1561 Flexser, A. J. and Tulving, E. (1982). Priming and recognition failure. *Journal of Verbal*  
1562 *Learning and Verbal Behavior*, 21:237–248.
- 1563 Flores, S., Bailey, H. R., Eisenberg, M. L., and Zacks, J. M. (2017). Event segmentation  
1564 improves event memory up to one month later. *Journal of Experimental Psychology:*  
1565 *Learning, Memory, and Cognition*, 43(8):1183.
- 1566 Gershman, S. J., Schapiro, A. C., Hupbach, A., and Norman, K. A. (2013). Neural context  
1567 reinstatement predicts memory misattribution. *The Journal of Neuroscience*, 33(20):8590–  
1568 8595.
- 1569 Glenberg, A. M., Bradley, M. M., Kraus, T. A., and Renzaglia, G. J. (1983). Studies of the  
1570 long-term recency effect: support for a contextually guided retrieval theory. *Journal of*  
1571 *Experimental Psychology: Learning, Memory, and Cognition*, 12:413–418.
- 1572 Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather  
1573 prediction” task? individual variability in strategies for probabilistic category learning.  
1574 *Learning and Memory*, 9:408–418.
- 1575 Gold, D. A., Zacks, J. M., and Flores, S. (2017). Effects of cues to event segmentation on  
1576 subsequent memory. *Cognitive Research: Principles and Implications*, 2(1):1.
- 1577 Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder,  
1578 A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto,  
1579 C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A.,  
1580 Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. (2021). Thinking

1581 ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*,  
1582 page doi.org/10.1101/2020.12.02.403477.

1583 Gotts, S. J., Chow, C. C., and Martin, A. (2012). Repetition priming and repetition sup-  
1584 pression: A case for enhanced efficiency through neural synchronization. *Cognitive*  
1585 *Neuroscience*, 3(3-4):227–237.

1586 Griffiths, T. L. and Steyvers, M. (2003). Prediction and semantic association. *Advances in*  
1587 *Neural Information Processing Systems*, 15.

1588 Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml,  
1589 M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages  
1590 2338–2342.

1591 Hargreaves, I. S., Pexman, P. M., Johnson, J. C., and Zdrazilova, L. (2012). Richer concepts  
1592 are better remembered: number of features effects in free recall. *Frontiers in Human*  
1593 *Neuroscience*, 6:doi.org/10.3389/fnhum.2012.00073.

1594 Healey, M. K. and Uitvlugt, M. G. (2019). The role of control processes in temporal and  
1595 semantic contiguity. *Memory and Cognition*, 47:719–737.

1596 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail:  
1597 a Python toolbox for analyzing and plotting free recall data. *Journal of Open Source*  
1598 *Software*, 10.21105/joss.00424.

1599 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal  
1600 behavioral and neural signatures of transforming experiences into memories. *Nature*  
1601 *Human Behavior*, 5:905–919.

1602 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). HyperTools: a

- 1603 Python toolbox for gaining geometric insights into high-dimensional data. *Journal of*  
1604 *Machine Learning Research*, 18(152):1–6.
- 1605 Hogan, R. M. (1975). Interitem encoding and directed search in free recall. *Memory and*  
1606 *Cognition*, 3:197–209.
- 1607 Howard, M. W. and Kahana, M. J. (1999). Contextual variability and serial position effects  
1608 in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:923–  
1609 941.
- 1610 Howard, M. W. and Kahana, M. J. (2002a). A distributed representation of temporal  
1611 context. *Journal of Mathematical Psychology*, 46:269–299.
- 1612 Howard, M. W. and Kahana, M. J. (2002b). When does semantic similarity help episodic  
1613 retrieval? *Journal of Memory and Language*, 46:85–98.
- 1614 Huang, L., Holcombe, A. O., and Pashler, H. (2004). Repetition priming in visual search:  
1615 episodic retrieval, not feature priming. *Memory and Cognition*, 32:12–20.
- 1616 Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental*  
1617 *Psychology: General*, 137(2):324–347.
- 1618 Huber, D. E., Shiffrin, R. M., Lyle, K. B., and Ruys, K. I. (2001). Perception and preference  
1619 in short-term word priming. *Psychological Review*, 108(1):149–182.
- 1620 Isarida, T. and Isarida, T. K. (2007). Environmental context effects of background color in  
1621 free recall. *Memory and Cognition*, 35(7):1620–1629.
- 1622 Jenkins, J. J. and Russell, W. A. (1952). Associative clustering during recall. *Journal of*  
1623 *Abnormal and Social Psychology*, 47:818–821.



- 1624 Jones, A. C. and Pyc, M. A. (2014). The production effect: costs and benefits in free recall.  
1625 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):300–305.
- 1626 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing  
1627 prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 1628 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*,  
1629 24:103–109.
- 1630 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York,  
1631 NY.
- 1632 Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychol-*  
1633 *ogy*, 71:107–138.
- 1634 Kahana, M. J., Howard, M. W., and Polyn, S. M. (2008). Associative processes in episodic  
1635 memory. In Roediger III, H. L., editor, *Cognitive Psychology of Memory*, pages 476–490.  
1636 Elsevier, Oxford, UK.
- 1637 Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., and Yeshurun, Y. (2023). Deeper than  
1638 you think: partisanship-dependent brain responses in early sensory and motor brain  
1639 regions. *The Journal of Neuroscience*, pages doi.org/10.1523/JNEUROSCI.0895–22.2022.
- 1640 Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning  
1641 of memories by context-based prediction error. *Proceedings of the National Academy of*  
1642 *Sciences, USA*, In press.
- 1643 Kimball, D. R., Smith, T. A., and Kahana, M. J. (2007). The fSAM model of false recall.  
1644 *Psychological Review*, 114(4):954–993.
- 1645 Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.

- 1646 Lange, E. B. (2005). Disruption of attention by irrelevant stimuli in serial recall. *Journal of*  
1647 *Memory and Language*, 43(4):513–531.
- 1648 Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases  
1649 of human memory. *Current Opinion in Behavioral Sciences*, 32:111–119.
- 1650 Lohnas, L. J., Polyn, S. M., and Kahana, M. J. (2010). Modeling intralist and interlist effects  
1651 in free recall. In *Psychonomic Society*, Saint Louis, MO.
- 1652 Luek, S. P., McLaughlin, J. P., and Cicala, G. A. (1971). Effects of blocking of input and  
1653 blocking of retrieval cues on free recall learning. *Journal of Experimental Psychology*,  
1654 91(1):159–161.
- 1655 Madan, C. R. (2021). Exploring word memorability: how well do different word properties  
1656 explain item free-recall probability? *Psychonomic Bulletin and Review*, 28:583–595.
- 1657 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
1658 *Handbook of Human Memory*. Oxford University Press.
- 1659 Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”  
1660 function? *Psychological Review*, 128(4):711–725.
- 1661 Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. A.  
1662 (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic*  
1663 *Bulletin and Review*, 23(5):1534–1542.
- 1664 Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free  
1665 recall. *Memory*, 20(5):511–517.
- 1666 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic  
1667 memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.

- 1668 Manning, J. R., Notaro, G. M., Chen, E., and Fitzpatrick, P. C. (2022). Fitness tracking  
1669 reveals task-specific associations between memory, mental health, and physical activity.  
1670 *Scientific Reports*, 12(13822):doi.org/10.1038/s41598-022-17781-0.
- 1671 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory pat-  
1672 terns in temporal lobe reveal context reinstatement during memory search. *Proceedings*  
1673 *of the National Academy of Sciences, USA*, 108(31):12893–12897.
- 1674 Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., and Kahana, M. J. (2012).  
1675 Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clus-  
1676 tering during memory search. *The Journal of Neuroscience*, 32(26):8871–8878.
- 1677 Marsh, J. E., Beaman, C. P., Hughes, R. W., and Jones, D. M. (2012). Inhibitory control in  
1678 memory: evidence for negative priming in free recall. *Journal of Experimental Psychology:*  
1679 *Learning, Memory, and Cognition*, 38(5):1377–1388.
- 1680 Marsh, J. E., Sörqvist, P., Hodgetts, H. M., Beaman, C. P., and Jones, D. M. (2015). Distraction  
1681 control processes in free recall: benefits and costs to performance. *Journal of Experimental*  
1682 *Psychology: Learning, Memory, and Cognition*, 41(1):118–133.
- 1683 Masicampo, E. J. and Sahakyan, L. (2014). Imagining another context during encoding off-  
1684 sets context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory,*  
1685 *and Cognition*, 40(6):1772–1777.
- 1686 Masís-Obando, R., Norman, K. A., and Baldassano, C. (2022). Scheme representations in  
1687 distinct brain networks support narrative memory during encoding and retrieval. *eLife*,  
1688 11:e70445.
- 1689 McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental*  
1690 *Psychology: Learning, Memory, and Cognition*, 20:507–520.

- 1691 Meinhardt, M. J., Bell, R., Buchner, A., and Röer, J. P. (2020). Adaptive memory: is  
1692 the animacy effect on memory due to richness of encoding? *Journal of Experimental*  
1693 *Psychology: Learning, Memory, and Cognition*, 46(3):416–426.
- 1694 Miller, J. F., Kahana, M. J., and Weidemann, C. T. (2012). Recall termination in free recall.  
1695 *Memory and Cognition*, 40(4):540–550.
- 1696 Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman,  
1697 S. J. (2017). The successor representation in human reinforcement learning. *Nature*  
1698 *Human Behavior*, 1:680–692.
- 1699 Moran, R. and Goshen-Gottstein, Y. (2014). The conditional-recency dissociation is con-  
1700 founded with nominal recency: should unitary models of memory still be devaluated?  
1701 *Psychonomic Bulletin and Review*, 21:332–343.
- 1702 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental*  
1703 *Psychology: General*, 64:482–488.
- 1704 Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy  
1705 of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1706 Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhi-  
1707 bitionless spreading activation and limited-capacity attention. *Journal of Experimental*  
1708 *Psychology: General*, 106(3):226–254.
- 1709 Oberauer, K. and Lewandowsky, S. (2008). Forgetting in immediate serial recall: decay,  
1710 temporal distinctiveness, or interference? *Psychological Review*, 115(3):544–576.
- 1711 Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., and Radvansky, G. A.  
1712 (2016). Event boundaries and memory improvement. *Cognition*, 148:136–144.

- 1713 Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of  
1714 context. *Trends in Cognitive Sciences*, 12:24–30.
- 1715 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). Task context and organization in  
1716 free recall. *Neuropsychologia*, 47:2158–2163.
- 1717 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly*  
1718 *Journal of Experimental Psychology*, 17:132–138.
- 1719 Puff, C. R. (1974). A consolidated theoretical view of stimulus-list organization effects in  
1720 free recall. *Psychological Reports*, 34:275–288.
- 1721 Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of  
1722 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,  
1723 NY.  
1724
- 1725 Rabinowitz, J. C. (1986). Priming in episodic memory. *Journal of Gerontology*, 41:204–213.
- 1726 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:  
1727 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.
- 1728 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition.  
1729 *Current Opinion in Behavioral Sciences*, 17:133–140.
- 1730 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior.  
1731 *Nature Reviews Neuroscience*, 13:713–726.
- 1732 Reinitz, M. T., Lammers, W. J., and Cochran, B. P. (1992). Memory-conjunction errors:  
1733 miscombination of stored stimulus features can produce illusions of memory. *Memory and Cognition*, 20:1–11.  
1734

- 1735 Rissman, J., Eliassen, J. C., and Blumstein, S. E. (2003). An event-related fMRI investigation  
1736 of implicit semantic priming. *Journal of Cognitive Neuroscience*, 15(8):1160–1175.
- 1737 Romney, A. K., Brewer, D. D., and Batchelder, W. H. (1993). Predicting clustering from  
1738 semantic structure. *Psychological Science*, 4:28–34.
- 1739 Sahakyan, L. and Kelley, C. M. (2002). A contextual change account of the directed  
1740 forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,  
1741 28(6):1064–1072.
- 1742 Sahakyan, L. and Smith, J. R. (2014). A long time ago, in a context far, far away: Retro-  
1743 spective time estimates and internal context change. *Journal of Experimental Psychology:*  
1744 *Learning, Memory, and Cognition*, 40(1):86–93.
- 1745 Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclo-*  
1746 *pedic Reference*, 3:501–506.
- 1747 Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008). A context-based theory of  
1748 recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.
- 1749 Sederberg, P. B., Miller, J. F., Howard, W. H., and Kahana, M. J. (2010). The tempo-  
1750 ral contiguity effect predicts episodic memory performance. *Memory and Cognition*,  
1751 38(6):689–699.
- 1752 Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of  
1753 time. *Neural Computation*, 24:134–193.
- 1754 Shapiro, S. I. (1970). Isolation effects, free recall, and organization. *Journal of Psychology*,  
1755 24:178–183.

- 1756 Sirotin, Y. B., Kimball, D. R., and Kahana, M. J. (2005). Going beyond a single list: modeling  
1757 the effects of prior experience on episodic free recall. *Psychonomic Bulletin and Review*,  
1758 12(5):787–805.
- 1759 Slamecka, N. J. and Barlow, W. (1979). The role of semantic and surface features in word  
1760 repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 18:617–627.
- 1761 Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: a review and  
1762 meta-analysis. *Psychonomic Bulletin and Review*, 8(2):203–220.
- 1763 Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009). A  
1764 Bayesian analysis of dynamics in free recall. *Advances in Neural Information Processing*  
1765 *Systems*, 22.
- 1766 Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., and Zacks, J. M. (2011).  
1767 Changes in events alter how people remember recent information. *Journal of Cognitive*  
1768 *Neuroscience*, 23(5):1052–1064.
- 1769 Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception  
1770 affect memory encoding and updating. *Journal of Experimental Psychology: General*,  
1771 138(2):236–257.
- 1772 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in*  
1773 *Cognitive Sciences*, 22(3):201–212.
- 1774 Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *The*  
1775 *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37:571–  
1776 590.
- 1777 Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P.,

- 1778 and Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, 316(5821):76–  
1779 82.
- 1780 Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press, New York, NY.
- 1781 Tulving, E. and Schacter, D. L. (1991). Priming and human memory systems. *Science*,  
1782 247:301–305.
- 1783 Watkins, P. C., Mathews, A., Williamson, D. A., and Fuller, R. D. (1992). Mood-congruent  
1784 memory in depression: emotional priming or elaboration? *Journal of Abnormal Psychol-*  
1785 *ogy*, 101(3):581–586.
- 1786 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American*  
1787 *Journal of Psychology*, 35:396–401.
- 1788 Whitely, P. L. (1927). The dependence of learning and recall upon prior intellectual activi-  
1789 ties. *Journal of Experimental Psychology: General*, 10:489–508.
- 1790 Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming.  
1791 *Current Opinion in Neurobiology*, 8(2):227–233.
- 1792 Xu, X., Zhu, Z., and Manning, J. R. (2023). The psychological arrow of time drives  
1793 temporal asymmetries in retrodicting versus predicting narrative events. *PsyArXiv*,  
1794 page doi.org/10.31234/osf.io/yp2qu.
- 1795 Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., and Hasson, U.  
1796 (2017). Same story, different story: the neural representation of interpretive frameworks.  
1797 *Psychological Science*, 28(3):307–319.
- 1798 Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., and



- 1799 Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal*  
1800 *of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792–804.
- 1801 Zhang, Q., Griffiths, T. L., and Norman, K. A. (2023). Optimal policies for free recall.  
1802 *Psychological Review*, 130(4):1104–1125.
- 1803 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018).  
1804 Is automatic speech-to-text transcription ready for use in psychological experiments?  
1805 *Behavior Research Methods*, 50:2597–2605.
- 1806 Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation  
1807 models in narrative comprehension: an event-indexing model. *Psychological Science*,  
1808 6(5):292–297.
- 1809 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension  
1810 and memory. *Psychological Bulletin*, 123(2):162–185.