

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

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

¹Dartmouth College

²Princeton University

³Akili Interactive

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

4 **Abstract**

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

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

19 Introduction

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

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

33 Our understanding of how we form situation models and schemas, and how they
34 interact with our subjective experiences and memories, is constrained in part by substantial
35 differences in how we study these processes. Situation models and schemas are most often
36 studied using “naturalistic” stimuli such as narratives and movies (Nastase et al., 2020;
37 Zwaan et al., 1995; Zwaan and Radvansky, 1998). In contrast, our understanding of how
38 we organize our memories has been most widely informed by more traditional paradigms
39 like free recall of random word lists (Kahana, 2012, 2020). In free recall, participants study
40 lists of items and are instructed to recall the items in any order they choose. The orders
41 in which words come to mind can provide insights into how participants have organized
42 their memories of the studied words. Because random word lists are unstructured by
43 design, it is not clear if or how non-trivial situation models might apply to these stimuli.

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

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

58 Over the past half-century, context-based models have enjoyed impressive success at
59 explaining many stereotyped behaviors observed during free recall and other list-learning
60 tasks (Estes, 1955; Glenberg et al., 1983; Howard and Kahana, 2002; Kimball et al., 2007;
61 Polyn and Kahana, 2008; Polyn et al., 2009; Raaijmakers and Shiffrin, 1980; Sederberg et al.,
62 2008; Shankar and Howard, 2012; Sirotin et al., 2005). These phenomena include the well-
63 known recency and primacy effects (superior recall of items from the end and, to a lesser
64 extent, from the beginning of the study list), as well as semantic and temporal clustering
65 effects (Kahana et al., 2008). The contiguity effect is an example of temporal clustering,
66 which is perhaps the dominant form of organization in free recall. This effect can be
67 seen in the tendency for people to successively recall items that occupied neighboring
68 positions in the studied list (Kahana, 1996). There are also striking effects of semantic

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

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

92 We designed an experimental paradigm to explore how people organize their mem-
93 ories for simple stimuli (word lists) whose temporal properties change across different

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

114 **Materials and methods**

115 **Participants**

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

122 Participants received course credit for enrolling in our study. We asked each partic-
123 ipant to fill out a demographic survey that included questions about their age, gender,
124 ethnicity, race, education, vision, reading impairments, medications or recent injuries,
125 coffee consumption on the day of testing, and level of alertness at the time of testing. All
126 components of the demographics survey were optional. One participant elected not to fill
127 out any part of the demographic survey, and all other participants answered some or all
128 of the survey questions.

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

138 exceed our target enrollments for several conditions. Nevertheless, we analyze all viable
139 data in the present paper.

140 Participants were assigned to experimental conditions based loosely on their date of
141 participation. (This aspect of our procedure helped us to more easily synchronize the ex-
142 periment databases across multiple testing computers.) Of the 490 participants who opted
143 to fill out the demographics survey, reported ages ranged from 17 to 31 years (mean: 19.1
144 years; standard deviation: 1.356 years). A total of 318 participants reported their gender as
145 female, 170 as male, and two participants declined to report their gender. A total of 442 par-
146 ticipants reported their ethnicity as “not Hispanic or Latino,” 39 as “Hispanic or Latino,”
147 and nine declined to report their ethnicity. Participants reported their races as White (345
148 participants), Asian (120 participants), Black or African American (31 participants), Amer-
149 ican Indian or Alaska Native (11 participants), Native Hawaiian or Other Pacific Islander
150 (four participants), Mixed race (three participants), Middle Eastern (one participant), and
151 Arab (one participant). A total of five participants declined to report their race. We note
152 that several participants reported more than one of the above racial categories. Participants
153 reported their highest degrees achieved as “Some college” (359 participants), “High school
154 graduate” (117 participants), “College graduate” (seven participants), “Some high school”
155 (five participants), “Doctorate” (one participant), and “Master’s degree” (one participant).
156 A total of 482 participants reported no reading impairments, and eight reported having
157 mild reading impairments. A total of 489 participants reported having normal color vision
158 and one participant reported that they were red-green color blind. A total of 482 partic-
159 ipants reported taking no prescription medications and having no recent injuries; four
160 participants reported having ADHD, one reported having dyslexia, one reported having
161 allergies, one reported a recently torn ACL/MCL, and one reported a concussion from
162 several months prior. The participants reported consuming 0 – 3 cups of coffee prior to

163 the testing session (mean: 0.32 cups; standard deviation: 0.58 cups). Participants reported
164 their current level of alertness, and we converted their responses to numerical scores as
165 follows: “very sluggish” (-2), “a little sluggish” (-1), “neutral” (0), “a little alert” (1), and
166 “very alert” (2). Across all participants, the full range of alertness levels were reported
167 (range: -2 – 2; mean: 0.35; standard deviation: 0.89).

168 We dropped from our dataset the one participant who reported having abnormal color
169 vision, as well as 39 participants whose data were corrupted due to technical failures while
170 running the experiment or during the daily database merges. In total, this left usable data
171 from 452 participants, broken down by experimental condition as follows: feature rich (67
172 participants), reduced (61 participants), reduced (early), (42 participants), reduced (late)
173 (41 participants), category (30 participants), size (30 participants), length (30 participants),
174 first letter (30 participants), color (31 participants), location (30 participants), and adaptive
175 (60 participants). The participant who declined to fill out their demographic survey
176 participated in the location condition, and we verified verbally that they had normal color
177 vision and no significant reading impairments.

178 **Experimental design**

179 Our experiment is a variant of the classic free recall paradigm that we term *feature-rich free*
180 *recall*. In feature-rich free recall, participants study 16 lists, each comprised of 16 words that
181 vary along a number of stimulus dimensions (Fig. 1). The stimulus dimensions include
182 two semantic features related to the *meanings* of the words (semantic category, referent
183 object size), two lexicographic features related to the *letters* that make up the words (word
184 length in number of letters, identity of the word’s first letter), and two visual features
185 that are independent of the words themselves (text color, presentation location). Each list
186 contains four words from each of four different semantic categories and two object sizes; all

187 other stimulus features are randomized. After studying each list, the participant attempts
188 to recall as many words as they can from that list, in any order they choose. Because
189 each individual word is associated with several well-defined (and quantifiable) features,
190 and because each list incorporates a diverse mix of feature values along each dimension,
191 this allows us to estimate which features participants are considering or leveraging in
192 organizing their memories.

193 **Stimuli**

194 The stimuli in our paradigm were 256 English words selected in a previous study (Ziman
195 et al., 2018). The words all referred to concrete nouns, and were chosen from 15 unique se-
196 mantic categories: body parts, building-related, cities, clothing, countries, flowers, fruits,
197 insects, instruments, kitchen-related, mammals, (US) states, tools, trees, and vegetables.
198 We also tagged each word according to the approximate size of the object the word re-
199 ferred to. Words were labeled as “small” if the corresponding object was likely able to
200 “fit in a standard shoebox” or “large” if the object was larger than a shoebox. Semantic
201 categories varied in how many object sizes they reflected (mean number of different sizes
202 per category: 1.33; standard deviation: 0.49). The numbers of words in each semantic
203 category also varied from 12 – 28 (mean number of words per category: 17.07; standard
204 deviation number of words: 4.65). We also identified lexicographic features for each word,
205 including the words’ first letters and lengths (i.e., number of letters). Across all categories,
206 all possible first letters were represented except for ‘Q’ (average number of unique first
207 letters per category: 11; standard deviation: 2 letters). Word lengths ranged from 3 – 12
208 letters (average: 6.17 letters; standard deviation: 2.06 letters).

209 We assigned the categorized words into a total of 16 lists with several constraints. First,
210 we required that each list contained words from exactly four unique categories, each with



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 lists participants might encounter in each condition. The rectangles during the “Presentation phase” show illustrated screen captures during a series of word presentations. Each word appeared onscreen for 2 seconds, followed by 2 seconds of blank screen. The red microphone icons during the “Recall” phase denote the one minute verbal recall interval. The labels on the right (and corresponding groupings on the left) denote experimental condition labels.

211 exactly four exemplars from each category. Second, we required that (across all words
212 on the list) at least one instance of both object sizes were represented. On average, each
213 category was represented in 4.27 lists (standard deviation: 1.16 lists). Aside from these
214 two constraints, we assigned each word to a unique list. After random assignment, each
215 list contained words with an average of 11.13 unique starting letters (standard deviation:
216 1.15 letters) and an average word length of 6.17 letters (standard deviation: 0.34 letters).

217 The above assignments of words to lists was performed once across all participants,
218 such that every participant studied the same set of 16 lists. In every condition we random-
219 ized the study order of these lists across participants. For participants in some conditions,
220 on some lists, we also randomly varied two additional visual features associated with each
221 word: the presentation font color, and the word’s onscreen location. These attributes were
222 assigned independently for each word (and for every participant). These visual features
223 were varied for words in all lists and conditions except for the “reduced” condition (all
224 lists), the first eight lists of the “reduced (early)” condition, and the last eight lists of the
225 “reduced (late)” condition. In these latter cases, words were all presented in black at the
226 center of the experimental computer’s display.

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

235 Most of the experimental manipulations we carried out entailed presenting or sorting

the presented words differently on the first eight lists participants studied (which we call *early* lists) versus on the final eight lists they studied (*late* lists). Since every participant studied exactly 16 lists, every list was either “early” or “late” depending on its order in the list study sequence.

Real-time speech-to-text processing

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

Random conditions (Fig. 1, top four rows)

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

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

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

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

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

are *discrete*. For these features we defined a discrete distance function. Specifically, we defined the distance between two words as either the absolute difference between their lengths, or the absolute distance between their starting letters in the English alphabet, respectively. For example, two words that started with the same letter would have a “first letter” distance of 0, and words starting with ‘J’ and ‘A’ respectively would have a first letter distance of 9. Because words’ lengths and letters’ positions in the alphabet are always integers, these discrete distances always take on integer values. Finally, the visual features (color and location) are *continuous* and *multivariate*, in that each “feature” takes on multiple (positive) real values. We defined the “color” and “location” distances between two words as the Euclidean distances between their (r, g, b) color or (x, y) location vectors, respectively. Therefore the color and location distance measures always take on non-negative real values (upper-bounded at 441.67 for color, or 27 in for location, reflecting the distances between the corresponding maximally different vectors).

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

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

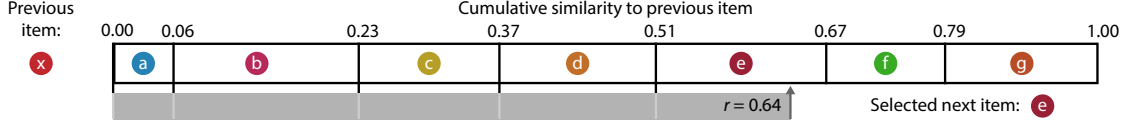


Figure 2: Generating stochastic feature-sorted lists. For a given feature dimension (e.g., color), we compute the similarity (Eqn. 1) between the feature value(s) of the previous item, x , and all yet-to-be-presented items ($a - g$). Next, we normalize these similarity scores so that they sum to 1. We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

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

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

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

310 As illustrated in Figure 2, we use these normalized similarity scores to construct a
 311 sequence of “sticks” that we lay end to end in a line. Each of the n sticks corresponds
 312 to a single to-be-presented word, and the stick lengths are proportional to the relative
 313 similarities between each word’s feature value(s) and the feature value(s) of the just-
 314 presented word. We choose the next to-be-presented word by moving an indicator along
 315 the set of sticks, by a distance chosen uniformly and at random on the interval $[0, 1]$. We
 316 select the word associated with the stick lying next to the indicator to be presented next.
 317 This process continues iteratively (re-computing the similarity scores and stochastically

318 choosing the next to-be-presented word using the just-presented word) until all of the
319 words have been presented. The result is an ordered list that tends to change gradually
320 along the selected feature dimension.

321 **Adaptive condition**

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

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

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

340 Lists in the random batches were sorted randomly (as on the initialization lists and in
341 the feature rich condition). Lists in the stabilize and destabilize batches were sorted in ways

342 that either matched or mismatched each participant’s memory fingerprint, respectively.
343 Our procedures for estimating participants’ memory fingerprints and ordering the stabilize
344 and destabilize lists are described next.

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

358 **Temporal clustering score (uncorrected).** Temporal clustering describes a participant’s
359 tendency to organize their recall sequences by the learned items’ encoding positions. For
360 instance, if a participant recalled the lists’ words in the exact order they were presented (or
361 in exact reverse order), this would yield a score of 1. If a participant recalled the words in
362 a random order, this would yield an expected score of 0.5. For each recall transition (and
363 separately for each participant), we sorted all not-yet-recalled words according to their
364 absolute lag (that is, distance away in the list). We then computed the percentile rank of
365 the next word the participant recalled. We took an average of these percentile ranks across

all of the participant’s recalls to obtain a single (uncorrected) temporal clustering score for the participant.

Permutation-corrected feature clustering scores. Suppose that two lists contain unequal numbers of items of each size. For example, suppose that list *A* contains all “large” items, whereas list *B* contains an equal mix of “large” and “small” items. For a participant recalling list *A*, any correctly recalled item will necessarily match the size of the previous correctly recalled item. In other words, successively recalling several list *A* items of the same size is essentially meaningless, since *any* correctly recalled list *A* word will be large. In contrast, successively recalling several list *B* items of the same size *could* be meaningful, since (early in the recall sequence) the yet-to-be-recalled items come from a mix of sizes. However, once all of the small items on list *B* have been recalled, the best possible next matching recall will be a large item. And all subsequent correct recalls must also be large items— so for those later recalls it becomes difficult to determine whether the participant is successively recalling large items because they are organizing their memories according to size, or (alternatively), whether they are simply recalling the yet-to-be-recalled items in a random order. In general, the precise order and blend of feature values expressed in a given list, the orders and numbers of correct recalls a participant makes, the number of intervening presentation positions between successive recalls, and so on, can all affect the range of clustering scores that are possible to observe for a given list. An uncorrected clustering score therefore conflates participants’ actual memory organization with other “nuisance” factors.

Following our prior work (Heusser et al., 2017), we used a permutation-based correction procedure to help isolate the behavioral aspects of clustering that we were most interested in. After computing the uncorrected clustering score (for the given list and observed recall sequence), we compute a “null” distribution of n additional clustering

391 scores after randomly shuffling the order of the recalled words (we use $n = 500$ in the
392 present study). This null distribution represents an approximation of the range of cluster-
393 ing scores one might expect to observe by “chance,” given that a hypothetical participant
394 was *not* truly clustering their recalls, but where the hypothetical participant still studied
395 and recalled exactly the same items (with the same features) as the true participant. We
396 define the *permutation-corrected clustering score* as the percentile rank of the observed un-
397 corrected clustering score in this estimated null distribution. In this way, a corrected score
398 of 1 indicates that the observed score was greater than any clustering score one might
399 expect by chance; in other words, good evidence that the participant was truly clustering
400 their recalls along the given feature dimension. We applied this correction procedure to
401 all of the clustering scores (feature and temporal) reported in this paper.

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

411 **Online “fingerprint” analysis.** The presentation orders of some lists in the adaptive
412 condition of our experiment (see *Adaptive condition*) were sorted according to participants’
413 *current* memory fingerprint, estimated using all of the lists they had studied up to that point
414 in the experiment. Because our experiment incorporated a speech-to-text component, all

415 of the behavioral data for each participant could be analyzed just a few seconds after the
416 conclusion of the recall intervals for each list. We used the Quail Python package (Heusser
417 et al., 2017) to apply speech-to-text algorithms to the just-collected data, aggregate the data
418 for the given participant, and estimate the participant’s memory fingerprint using all of
419 their available data up to that point in the experiment. Two aspects of our implementation
420 are worth noting. First, because memory fingerprints are computed independently for
421 each list and then averaged across lists, the already-computed memory fingerprints for
422 earlier lists could be cached and loaded as needed in future computations. This meant
423 that our computations pertaining to updating our estimate of a participant’s memory
424 fingerprint only needed to consider data from the most recent list. Second, each element
425 of the null distributions of uncorrected fingerprint scores (see *Permutation-corrected feature*
426 *clustering scores*) could be estimated independently from the others. This enabled us
427 to make use of the testing computers’ multi-core CPU architectures by considering (in
428 parallel) elements of the null distributions in batches of eight (i.e., the number of CPU
429 cores on each testing computer). Taken together, we were able to compress the relevant
430 computations into just a few seconds of computing time. The combined processing time for
431 the speech-to-text algorithm, fingerprint computations, and permutation-based ordering
432 procedure (described next) easily fit within the inter-list intervals, where participants
433 paused for a self-paced break before moving on to study and recall the next list.

434 **Ordering “stabilize” and “destabilize” lists by an estimated fingerprint.** In the adap-
435 tive condition of our experiment, the presentation orders for *stabilize* and *destabilize* lists
436 were chosen to either maximally or minimally (respectively) comport with participants’
437 memory fingerprints. Given a participant’s memory fingerprint and a to-be-presented set
438 of items, we designed a permutation-based procedure for ordering the items. First, we
439 dropped from the participant’s fingerprint the temporal clustering score. For the remain-

440 ing feature dimensions, we arranged the clustering scores in the fingerprint into a template
 441 vector, f . Second, we computed $n = 2500$ random permutations of the to-be-presented
 442 items. These permutations served as candidate presentation orders. We sought to select
 443 the specific order that most (or least) matched f . Third, for each random permutation, we
 444 computed the (permutation-corrected) “fingerprint,” treating the permutation as though
 445 it were a potential “perfect” recall sequence. (We did not include temporal clustering
 446 scores in these fingerprints.) This yielded a “simulated fingerprint” vector, \hat{f}_p for each
 447 permutation p . We used these simulated fingerprints to select a specific permutation, i ,
 448 that either maximized (for stabilize lists) or minimized (for destabilize lists) the correlation
 449 between \hat{f}_i and f .

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

451 Following some of our prior work (Heusser et al., 2021, 2018), we use low-dimensional
 452 embeddings to help visualize how participants’ memory fingerprints change across lists
 453 (Figs. 6A, S8A). To compute a shared embedding space across participants and experimen-
 454 tal conditions, we concatenated the full set of across-participant average fingerprints (for
 455 all lists and experimental conditions) to create a large matrix with number-of-lists ($16 \times$
 456 number-of-conditions (10, including the adaptive condition) rows and seven columns (one
 457 for each feature clustering score, plus an additional temporal clustering score column). We
 458 used principal components analysis to project the seven-dimensional observations into a
 459 two-dimensional space (using the two principal components that explained the most vari-
 460 ance in the data). For two visualizations (Figs. 6B, and S8B) we computed an additional
 461 set of two-dimensional embeddings for the *average* fingerprints across lists within a given
 462 list grouping (i.e., early or late). For those visualizations, we averaged across the rows (for
 463 each condition and group of lists) in the combined fingerprint matrix prior to projecting it

464 into the shared two-dimensional space. This yielded a single two-dimensional coordinate
465 for each *list group* (in each condition), rather than for each individual list. We used these
466 embeddings solely for visualization. All statistical tests were carried out in the original
467 (seven-dimensional) feature spaces.

468 **Analyses**

469 **Probability of n^{th} recall curves**

470 Probability of first recall curves (Atkinson and Shiffrin, 1968; Postman and Phillips, 1965;
471 Welch and Burnett, 1924) reflect the probability that an item will be recalled first, as a
472 function of its serial position during encoding. To carry out this analysis, we initialized
473 (for each participant) a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s.
474 Then, for each list, we found the index of the word that was recalled first, and we filled
475 in that position in the matrix with a 1. Finally, we averaged over the rows of the matrix
476 to obtain a 1 by 16 array of probabilities, for each participant. We used an analogous
477 procedure to compute probability of n^{th} recall curves for each participant. Specifically,
478 we filled in the corresponding matrices according to the n^{th} recall on each list that each
479 participant made. When a given participant had made fewer than n recalls for a given
480 list, we simply excluded that list from our analysis when computing that participant's
481 curve(s). The probability of first recall curve corresponds to a special case where $n = 1$.

482 **Lag-conditional response probability curve**

483 The lag-conditional probability (lag-CRP) curve (Kahana, 1996) reflects the probability of
484 recalling a given item after the just-recalled item, as a function of their relative encoding
485 positions (lag). In other words, a lag of 1 indicates that a recalled item was presented
486 immediately after the previously recalled item, and a lag of -3 indicates that a recalled item

came three items before the previously recalled item. For each recall transition (following the first recall), we computed the lag between the just-recalled word's presentation position and the next-recalled word's presentation position. We computed the proportions of transitions (between successively recalled words) for each lag, normalizing for the total numbers of possible transitions. In carrying out this analysis, we excluded all incorrect recalls and successive repetitions (e.g., recalling the same word twice in a row). This yielded, for each list, a 1 by number-of-lags (–15 to +15; 30 lags in total, excluding lags of 0) array of conditional probabilities. We averaged these probabilities across lists to obtain a single lag-CRP for each participant. Because transitions at large absolute lags are rare, these curves are typically displayed using range restrictions (Kahana, 2012).

Serial position curve

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

508 Identifying event boundaries

509 We used the distances between feature values for successively presented words (see *Defin-*
510 *ing feature-based distances*) to estimate “event boundaries” where the feature values changed
511 more than usual (DuBrow and Davachi, 2016; Ezzyat and Davachi, 2011; Manning et al.,
512 2016; Radvansky and Copeland, 2006; Swallow et al., 2011, 2009). For each list, for each
513 feature dimension, we computed the distribution of distances between the feature values
514 for successively presented words. We defined event boundaries (e.g., Fig. 3B) as occurring
515 between any successive pair of words whose distances along the given feature dimension
516 were greater than one standard deviation above the mean for that list. Note that, because
517 event boundaries are defined for each feature dimension, each individual list may contain
518 several sets of event boundaries, each at different moments in the presentation sequence
519 (depending on the feature dimension of interest).

520 Results

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

531 irrelevant information). Second, we manipulated the orders in which words were studied
532 (and how those orderings changed over time). We wondered whether presenting the same
533 list of words with different appearances (e.g., by manipulating font size and onscreen
534 location) or in different orders (e.g., sorted along one feature dimension versus another)
535 might serve to influence how participants organized their memories of the words. We also
536 wondered whether some order manipulations might be temporally “sticky” by influencing
537 how *future* lists were remembered.

538 To obtain a clean preliminary estimate of the consequences on memory of randomly
539 varying the font colors and locations of presented words (versus holding the font color
540 fixed at black, and holding the display locations fixed at the center of the display) we
541 compared participants’ performance on the *feature rich* and *reduced* experimental condi-
542 tions (see *Random conditions*, Fig. S1). In the feature rich condition the words’ colors and
543 locations varied randomly across words, and in the reduced condition words were always
544 presented in black, at the center of the display. Aggregating across all lists for each par-
545 ticipant, we found no difference in recall accuracy for feature rich versus reduced lists
546 ($t(126) = -0.290, p = 0.772$). However, participants in the feature rich condition clustered
547 their recalls substantially more along every dimension we examined (temporal clustering:
548 $t(126) = 10.624, p < 0.001$; category clustering: $t(126) = 10.077, p < 0.001$; size clustering:
549 $t(126) = 11.829, p < 0.001$; word length clustering: $t(126) = 10.639, p < 0.001$; first let-
550 ter clustering: $t(126) = 7.775, p < 0.001$; see *Permutation-corrected feature clustering scores*
551 for more information about how we quantified each participant’s clustering tendencies.)
552 Taken together, these comparisons suggest that adding new features changes how par-
553 ticipants organize their memories of studied words, even when those new features are
554 independent of the words themselves and even when the new features vary randomly
555 across words. We found no evidence that those additional uninformative features were

distracting (in terms of their impact on memory performance), but they did affect participants' recall dynamics (measured via their clustering scores).

We also wondered whether adding these irrelevant visual features to later lists (after the participants had already studied impoverished lists), or removing the visual features from later lists (after the participants had already studied visually diverse lists) might affect memory performance. In other words, we sought to test for potential effects of changing the “richness” of participants' experiences over time. All participants studied and recalled a total of 16 lists; we defined *early* lists as the first eight lists and *late* lists as the last eight lists each participant encountered. To help interpret our results, we compared participants' memories on early versus late lists in the above feature rich and reduced conditions. Participants in both conditions remembered more words on early versus late lists (feature rich: $t(66) = 4.553, p < 0.001$; reduced: $t(60) = 2.434, p = 0.018$). Participants in the feature rich (but not reduced) conditions exhibited more temporal clustering on early versus late lists (feature rich: $t(66) = 2.318, p = 0.024$; reduced: $t(60) = 0.929, p = 0.357$). And participants in both conditions exhibited more semantic (category and size) clustering on early versus late lists (feature rich, category: $t(66) = 3.805, p < 0.001$; feature rich, size: $t(66) = 2.190, p = 0.032$; reduced, category: $t(60) = 2.856, p = 0.006$; reduced, size: $t(60) = 2.947, p = 0.005$). Participants in the reduced (but not feature rich) conditions exhibited more lexicographic clustering on early versus late lists (feature rich, word length: $t(66) = 0.161, p = 0.872$; feature rich, first letter: $t(66) = 0.410, p = 0.683$; reduced, word length: $t(60) = 3.528, p = 0.001$; reduced, first letter: $t(60) = 2.275, p = 0.026$). Taken together, these comparisons suggest that even when the presence or absence of irrelevant visual features is stable across lists, participants still exhibit some differences in their performance and memory organization tendencies for early versus late lists.

With these differences in mind, we next compared participants' memories on early

581 versus late lists for two additional experimental conditions (see *Random conditions*, Fig. S1).
 582 In a *reduced (early)* condition, we held the irrelevant visual features constant on early lists,
 583 but allowed them to vary randomly on late lists. In a *reduced (late)* condition, we allowed
 584 the irrelevant visual features to vary randomly on early lists, but held them constant
 585 on late lists. Given our above findings that (a) participants tended to remember more
 586 words and exhibit stronger clustering effects on feature rich (versus reduced) lists, and (b)
 587 participants tended to remember more words and exhibit stronger clustering effects on
 588 early (versus late) lists, we expected these early versus late differences to be enhanced in the
 589 reduced (early) condition and diminished in the reduced (late) condition. However, to our
 590 surprise, participants in *neither* condition exhibited reliable early versus late differences in
 591 accuracy (reduced (early): $t(41) = 1.499, p = 0.141$; reduced (late): $t(40) = 1.462, p =$
 592 0.152), temporal clustering (reduced (early): $t(41) = 0.998, p = 0.324$; reduced (late):
 593 $t(40) = 1.099, p = 0.278$), nor feature-based clustering (reduced (early), category: $t(41) =$
 594 $0.753, p = 0.456$; reduced (early), size: $t(41) = 0.721, p = 0.475$; reduced (early), length:
 595 $t(41) = 0.493, p = 0.625$; reduced (early), first letter: $t(41) = 0.780, p = 0.440$; reduced (late),
 596 category: $t(40) = -0.086, p = 0.932$; reduced (late), size: $t(40) = 0.746, p = 0.460$; reduced
 597 (late), length: $t(40) = 1.476, p = 0.148$; reduced (late), first letter: $t(40) = 0.966, p = 0.340$).
 598 We hypothesized that adding or removing the irrelevant features was acting as a sort
 599 of “event boundary” between early and late lists. In prior work, we (and others) have
 600 found that memories formed just after event boundaries can be enhanced (e.g., due to less
 601 contextual interference between pre- and post-boundary items; Manning et al., 2016).
 602 We found that *adding* irrelevant visual features on later lists that had not been present
 603 on early lists (as in the reduced (early) condition) served to enhance recall performance
 604 relative to conditions where all lists had the same blends of features (accuracy for feature
 605 rich versus reduced (early): $t(107) = -2.230, p = 0.028$; reduced versus reduced (early):

$t(101) = -2.045, p = 0.043$; also see Fig. S3A). However, *subtracting* irrelevant visual features on later lists that *had* been present on early lists (as in the reduced (late) condition) did not appear to impact recall performance (accuracy for feature rich versus reduced (late): $t(106) = -0.638, p = 0.525$; reduced versus reduced (late): $t(100) = -0.407, p = 0.685$). These comparisons suggest that recall accuracy has a directional component (i.e., accuracy is affected differently by removing features later that had been present earlier versus adding features later that had *not* been present earlier). In contrast, we found that participants exhibited more temporal and feature-based clustering when we added irrelevant visual features to *any* lists (comparisons of clustering on feature rich versus reduced lists are reported above; temporal clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions: $ts \leq -9.780, ps < 0.001$; feature-based clustering in reduced versus reduced (early) and reduced versus reduced (late) conditions: $ts \leq -5.443, ps < 0.001$). Temporal and feature-based clustering were not reliably different in the feature rich, reduced (early), and reduced (late) conditions (temporal clustering in feature rich versus reduced (early) and feature rich versus reduced (late) conditions: $ts \geq -1.434, ps \geq 0.154$; feature-based clustering in feature rich versus reduced (early) and feature rich versus reduced (late) conditions: $ts \geq -1.359, ps > 0.177$).

Taken together, our findings thus far suggest that adding item features that change over time, even when they vary randomly and independently of the items, can enhance participants' overall memory performance and can also enhance temporal and feature-based clustering. To the extent that the number of item features that vary from moment to moment approximates the "richness" of participants' experiences, our findings suggest that participants remember "richer" stimuli better and organize richer stimuli more reliably in their memories. Next, we turn to examine the memory effects of varying the temporal ordering of different stimulus features while holding the features themselves constant. We

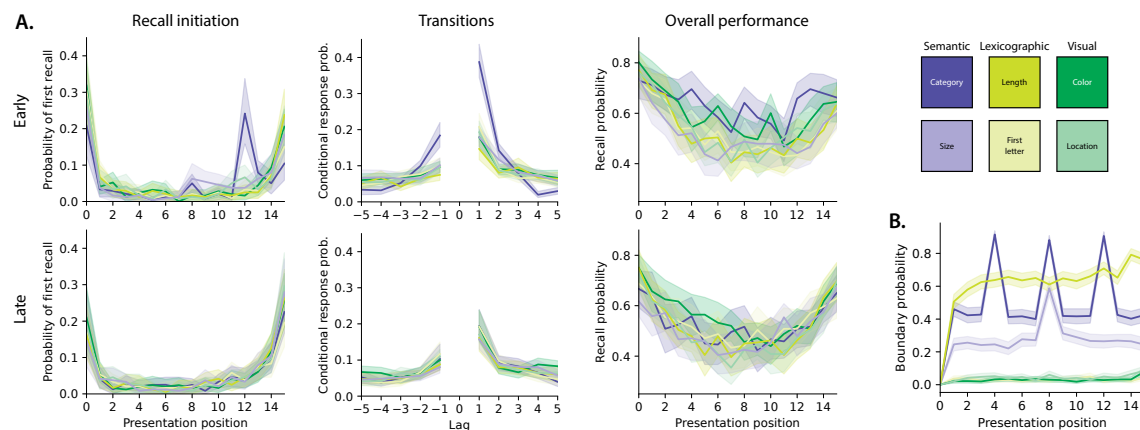


Figure 3: Recall dynamics in feature rich free recall (order manipulation conditions). **A.** Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling each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

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 condi-*

642 *tions*; features: category, size, length, first letter, color, and location). Participants in
 643 the category-ordered condition showed an increase in memory performance on early
 644 lists (accuracy, relative to early feature rich lists; $t(95) = 3.034, p = 0.003$). Partici-
 645 pants in the color-ordered condition also showed a trending increase in memory per-
 646 formance on early lists (again, relative to early feature rich lists: $t(96) = 1.850, p = 0.067$).
 647 Participants' performances on early lists in all of the other order manipulation con-
 648 ditions were indistinguishable from performance on the early feature rich lists ($|t|s$
 649 $< 1.013, ps > 0.314$). Participants in both of the semantically ordered conditions exhib-
 650 ited stronger temporal clustering on early lists (versus early feature rich lists; category:
 651 $t(95) = 8.508, p < 0.001$; size: $t(95) = 2.429, p = 0.017$). Participants in the length-ordered
 652 condition tended to exhibit *less* temporal clustering on early lists relative to early feature
 653 rich lists ($t(95) = -1.666, p = 0.099$), whereas participants in the first letter-ordered condi-
 654 tion exhibited stronger temporal clustering on early lists ($t(95) = 2.587, p = 0.011$). Partici-
 655 pants in the visually ordered conditions exhibited more similar performance on early lists,
 656 relative to early feature rich lists (color: $t(96) = -1.064, p = 0.290$; we found a trending
 657 enhancement for participants in the location-ordered condition: $t(95) = 1.682, p = 0.096$).
 658 We also compared feature-based clustering on early lists across the order manipulation
 659 and feature rich conditions. Since these results were similar across both semantic con-
 660 ditions (category and size), both lexicographic conditions (length and first letter), and
 661 both visual conditions (color and location), here we aggregate data from conditions that
 662 manipulated each of these three feature groupings in our comparisons, to simplify the
 663 presentation. On early lists, participants in the semantically ordered conditions exhibited
 664 stronger semantic clustering relative to participants in the feature rich condition (category:
 665 $t(125) = 2.524, p = 0.013$; size: $t(125) = 3.510, p = 0.001$), but showed no reliable differences
 666 in lexicographic (length: $t(125) = 0.539, p = 0.591$; first letter: $t(125) = -0.587, p = 0.558$)

667 or visual (color: $t(125) = -0.579, p = 0.564$; location: $t(125) = -0.346, p = 0.730$) clustering.
 668 Similarly, participants in the lexicographically ordered conditions exhibited stronger (rela-
 669 tive to feature rich participants) lexicographic clustering (length: $t(125) = 3.426, p = 0.001$;
 670 first letter: $t(125) = 3.236, p = 0.002$) on early lists, but showed no reliable differences in
 671 semantic (category: $t(125) = -1.078, p = 0.283$; size: $t(125) = -0.310, p = 0.757$) or visual
 672 (color: $t(125) = -0.209, p = 0.835$; location: $t(125) = -0.004, p = 0.997$) clustering. And
 673 participants in the visually ordered conditions exhibited stronger visual clustering (again,
 674 relative to feature rich participants, and on early lists; color: $t(126) = 2.099, p = 0.038$;
 675 location: $t(126) = 4.392, p < 0.001$), but showed now reliable differences in semantic (cat-
 676 egory: $t(126) = 0.204, p = 0.839$; size: $t(126) = -0.093, p = 0.926$) or lexicographic (length:
 677 $t(126) = 0.714, p = 0.476$; first letter: $t(126) = 0.820, p = 0.414$) clustering. Taken together,
 678 these order manipulation results suggest several broad patterns (Figs. 3A, 4). First, most of
 679 the order manipulations we carried out did *not* reliably affect overall recall performance.
 680 Second, most of the order manipulations increased participants' tendencies to temporally
 681 cluster their recalls. Third, all of the order manipulations enhanced participants' clus-
 682 tering of each condition's target feature (i.e., semantic manipulations enhanced semantic
 683 clustering, lexicographic manipulations enhanced lexicographic clustering, and visual
 684 manipulations enhanced visual clustering) while leaving clustering along other feature
 685 dimensions roughly unchanged (i.e., semantic manipulations did not affect lexicographic
 686 or visual clustering, and so on).

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

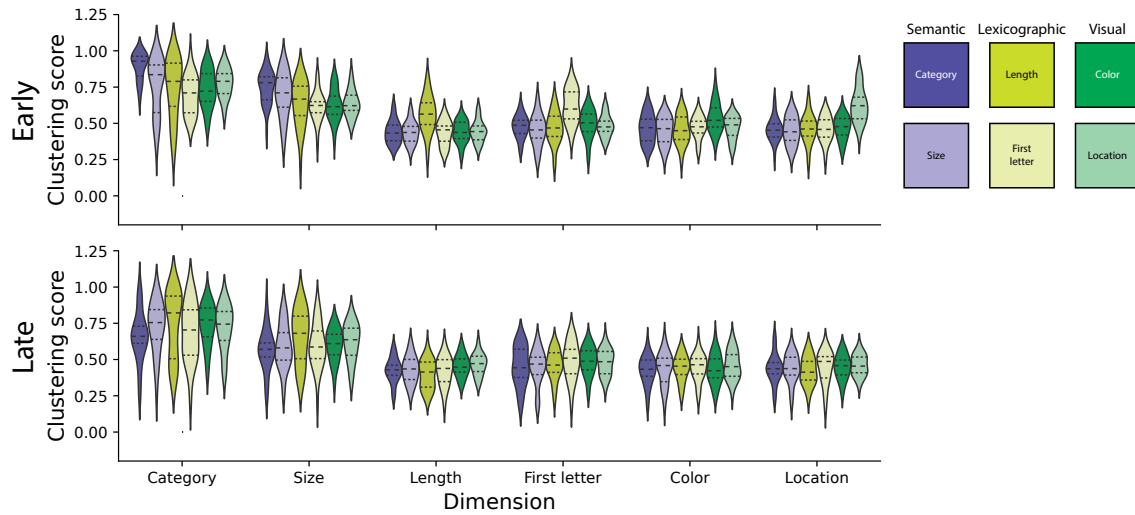


Figure 4: Memory “fingerprints” (order manipulation conditions). The across-participant distributions of clustering scores for each feature type (x-coordinate) are displayed for each experimental condition (color), separately for order manipulation (early, top) and randomly ordered (late, bottom) lists.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

692 who recalled randomly ordered lists tended to initiate recall with either the first or last list
 693 items (Fig. S1, top left panel). We hypothesized that the participants might be “clumping”
 694 their recalls into groups of items that shared category labels. Indeed, when we com-
 695 pared the positions of feature changes in the study sequence (Fig. 3B; see *Identifying event*
 696 *boundaries*) with the positions of items participants recalled first, we noticed a striking
 697 correspondence in both semantic conditions. Specifically, on category-ordered lists, the
 698 category labels changed every four items on average (dark purple peaks in Fig. 3B), and
 699 participants also seemed to display an increased tendency (relative to other order manipu-
 700 lation and random conditions) to initiate recall of category-ordered lists with items whose
 701 study positions were integer multiples of four. Similarly, for size-ordered lists, the size la-
 702 bels changed every eight items on average (light purple peaks in Fig. 3B), and participants
 703 also seemed to display an increased tendency to initiate recall of size-ordered lists with
 704 items whose study positions were integer multiples of eight. A second striking difference

705 is that participants in the category condition exhibited a much steeper lag-CRP (Fig. 3A,
706 top middle panel) than participants in other conditions. (This is another expression of
707 participants' increased tendencies to temporally cluster their recalls on category-ordered
708 lists, as we reported above.) Taken together, these order-specific idiosyncrasies suggest
709 a hierarchical set of influences on participants' memories. At longer timescales, "event
710 boundaries" (to use the term loosely) can be induced across lists by adding or removing
711 irrelevant visual features. At shorter timescales, "event boundaries" can be induced across
712 items (within a single list) by adjusting how item features change throughout the list.

713 The above comparisons between memory performance on early lists in the order ma-
714 nipulation versus feature rich conditions highlight how sorted lists are remembered differ-
715 ently from random lists. We also wondered how sorting lists along each feature dimension
716 influenced memory relative to sorting lists along the other feature dimensions. Partici-
717 pants trended towards remembering early lists that were sorted semantically better than
718 lexicographically sorted lists ($t(118) = 1.936, p = 0.055$). Participants also remembered
719 visually sorted lists better than lexicographically sorted lists ($t(119) = 2.145, p = 0.034$).
720 However, participants showed no reliable differences in recall for semantically versus
721 visually sorted lists ($t(119) = 0.113, p = 0.910$). Participants temporally clustered semanti-
722 cally sorted lists more strongly than either lexicographically ($t(118) = 5.572, p < 0.001$) or
723 visually ($t(119) = 6.215, p < 0.001$) sorted lists, but did not show reliable differences in tem-
724 poral clustering on lexicographically versus visually sorted lists ($t(119) = 0.189, p = 0.850$).
725 Participants also showed reliably more semantic clustering on semantically sorted lists
726 than lexicographically (category: $t(118) = 3.492, p = 0.001$, size: $t(118) = 3.972, p < 0.001$)
727 or visually (category: $t(119) = 2.702, p = 0.008$, size: $t(119) = 4.230, p < 0.001$) sorted
728 lists; more lexicographic clustering on lexicographically sorted lists than semantically
729 (length: $t(118) = 3.112, p = 0.002$; first letter: $t(118) = 3.686, p = 0.000$) or visually (length:

730 $t(119) = 3.024, p = 0.003$; first letter: $t(119) = 2.644, p = 0.009$) sorted lists; and more visual
731 clustering on visually sorted lists than semantically (color: $t(119) = -2.659, p = 0.009$;
732 location: $t(119) = -4.604, p < 0.001$) or lexicographically (color: $t(119) = -2.366, p = 0.020$;
733 location: $t(119) = -4.265, p < 0.001$) sorted lists. In summary, sorting lists by different
734 features appeared to have slightly different effects on overall memory performance and
735 temporal clustering, and people tended to cluster their recalls along a given feature di-
736 mension more when the studied lists were (versus were not) sorted along that dimension.

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

744 At the group level, there seemed to be almost no lingering impact of sorting early
745 lists on memory for later lists. To simplify the presentation, we report these null results
746 in aggregate across the three feature groupings. Relative to memory performance on
747 late feature rich lists, participants' memory performance in all six order manipulation
748 conditions showed no reliable differences (semantic: $t(125) = 0.487, p = 0.627$; lexico-
749 graphic: $t(125) = 0.878, p = 0.382$; visual: $t(126) = 1.437, p = 0.153$). Nor did we observe
750 any reliable differences in temporal clustering on late lists (relative to late feature rich
751 lists; semantic: $t(125) = 0.146, p = 0.884$; lexicographic: $t(125) = 0.923, p = 0.358$; visual:
752 $t(126) = 0.525, p = 0.601$). Aside from a slightly increased tendency for participants to
753 cluster words by their length on late visual order manipulation lists (more than late fea-
754 ture rich lists; $t(126) = 2.199, p = 0.030$), we observed no reliable differences in any type of

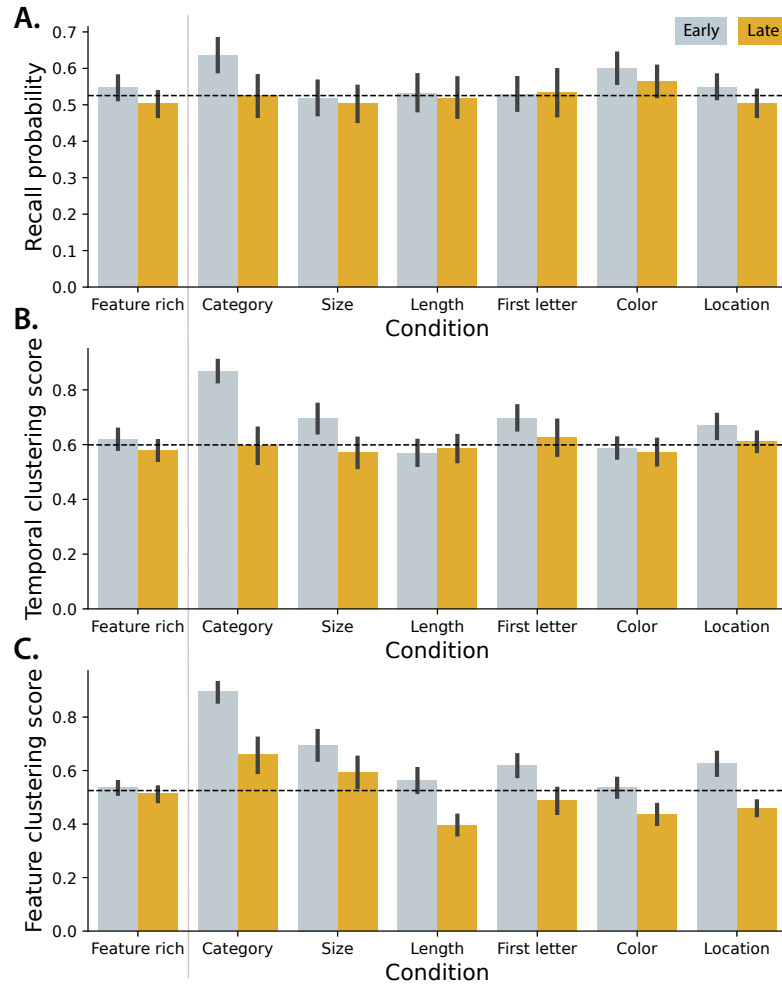


Figure 5: Recall probability and clustering scores on early and late lists. The bar heights display the average (across participants) recall probabilities (A.), temporal clustering scores (B.), and feature clustering scores (C.) for early (gray) and late (gold) lists. For the feature rich bars (left), the feature clustering scores are averaged across features. For the order manipulation conditions, feature clustering scores are displayed for the focused-on feature for each condition (e.g., category clustering scores are displayed for the category condition, and so on). All panels: error bars denote bootstrap-estimated 95% confidence intervals. The horizontal dotted lines denote the average values (across all lists and participants) for the feature rich condition.

755 feature clustering on late order manipulation condition lists versus late feature rich lists
756 ($\|t\|_s \leq 1.234, p_s \geq 0.220$).

757 We also looked for more subtle group-level patterns. For example, perhaps sorting
758 early lists by one feature dimension could affect how participants cluster *other* features (on
759 early and/or late lists) as well. We defined participants' *memory fingerprints* as the set of their
760 temporal and feature clustering scores. A participant's memory fingerprint describes how
761 they tend to retrieve memories of the studied items, perhaps searching through several
762 feature spaces (or along several representational dimensions). To gain insights into the
763 dynamics of how participants' clustering scores tended to change over time, we computed
764 the average (across participants) fingerprint from each list, from each order manipulation
765 condition (Fig. 6). We projected these fingerprints into a two-dimensional space to help
766 visualize the dynamics (top panels; see *Computing low-dimensional embeddings of memory*
767 *fingerprints*). We found that participants' average fingerprints tended to remain relatively
768 stable on early lists, and exhibited a "jump" to another stable state on later lists. The
769 sizes of these jumps varied somewhat across conditions (the Euclidean distances between
770 fingerprints in their original high dimensional spaces are displayed in the bottom panels).
771 We also averaged the fingerprints across early and late lists, respectively, for each condition
772 (Fig. 6B). We found that participants' fingerprints on early lists seem to be influenced by
773 the order manipulations for those lists (see the locations of the circles in Fig. 6B). There
774 also seemed to be some consistency across different features within a broader type. For
775 example, both semantic feature conditions (category and size; purple markers) diverge in
776 a similar direction from the group; both lexicographic feature conditions (length and first
777 letter; yellow markers) diverge in a similar direction; and both visual conditions (color
778 and location; green) also diverge in a similar direction. But on late lists, participants'
779 fingerprints seem to return to a common state that is roughly shared across conditions



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

780 (i.e., the stars in that panel are clumped together).

781 When we examined the data at the level of individual participants (Figs. 7 and 8), a
 782 clearer story emerged. Within each order manipulation condition, participants exhibited
 783 a range of feature clustering scores, on both early and late lists (Fig. 7A, B). Across every
 784 order manipulation condition, participants who exhibited stronger feature clustering (for
 785 their condition's manipulated feature) recalled more words. This trend held overall across
 786 conditions and participants (early: $r(179) = 0.537, p < 0.001$; late: $r(179) = 0.492, p < 0.001$)
 787 as well as for each condition individually for early ($r_s \geq 0.386$, all $p_s \leq 0.035$) and late
 788 ($r_s \geq 0.462$, all $p_s \leq 0.010$) lists. We found no evidence of a condition-level trend; for
 789 example the conditions where participants tended to show stronger clustering scores
 790 were not correlated with the conditions where participants remembered more words
 791 (early: $r(4) = 0.526, p = 0.284$; late: $r(4) = -0.257, p = 0.623$; see insets of panels A and
 792 B). We observed carryover associations between feature clustering and recall performance

(Fig. 7C, D). Participants who showed stronger feature clustering on early lists tended to recall more items on late lists (across conditions: $r(179) = 0.492, p < 0.001$; all conditions individually: $r_s \geq 0.462$, all $p_s \leq 0.010$). Participants who recalled more items on early lists also tended to show stronger feature clustering on late lists (across conditions: $r(179) = 0.280, p < 0.001$; all non-visual conditions: $r_s \geq 0.445$, all $p_s \leq 0.014$; color: $r(29) = 0.298, p = 0.103$; location: $r(28) = 0.354, p = 0.055$). Neither of these effects showed condition-level trends (early feature clustering versus late recall probability: $r(4) = -0.299, p = 0.565$; early recall probability versus late feature clustering: $r(4) = 0.400, p = 0.432$). We also looked for associations between feature clustering and temporal clustering. Across every order manipulation condition, participants who exhibited stronger feature clustering also exhibited stronger temporal clustering. For early lists (Fig. 7E), this trend held overall ($r(179) = 0.924, p < 0.001$), for each condition individually (all $r_s \geq 0.822$, all $p_s < 0.001$), and across conditions ($r(4) = 0.964, p = 0.002$). For late lists (Fig. 7F), the results were more variable (overall: $r(179) = 0.348, p < 0.001$; all non-visual conditions: $r_s \geq 0.382$, all $p_s \leq 0.037$; color: $r(29) = 0.453, p = 0.011$; location: $r(28) = 0.190, p = 0.314$; across-conditions: $r(4) = -0.036, p = 0.945$). While less robust than the carryover associations between feature clustering and recall performance, we also observed some carryover associations between feature clustering and temporal clustering (Fig. 7G, H). Participants who showed stronger feature clustering on early lists trended towards showing stronger temporal clustering on later lists (overall: $r(179) = 0.301, p < 0.001$; for individual conditions: all $r_s \geq 0.297$, all $p_s \leq 0.111$; across conditions: $r(4) = 0.107, p = 0.840$). And participants who showed stronger temporal clustering on early lists trended towards showing stronger feature clustering on later lists (overall: $r(179) = 0.579, p < 0.001$; all non-visual conditions: $r_s \geq 0.323$, all $p_s \leq 0.082$; visual conditions: $r_s \geq 0.089$, all $p_s \leq 0.632$; across conditions: $r(4) = 0.916, p = 0.010$). Taken together, the results displayed in Figure 7 show that

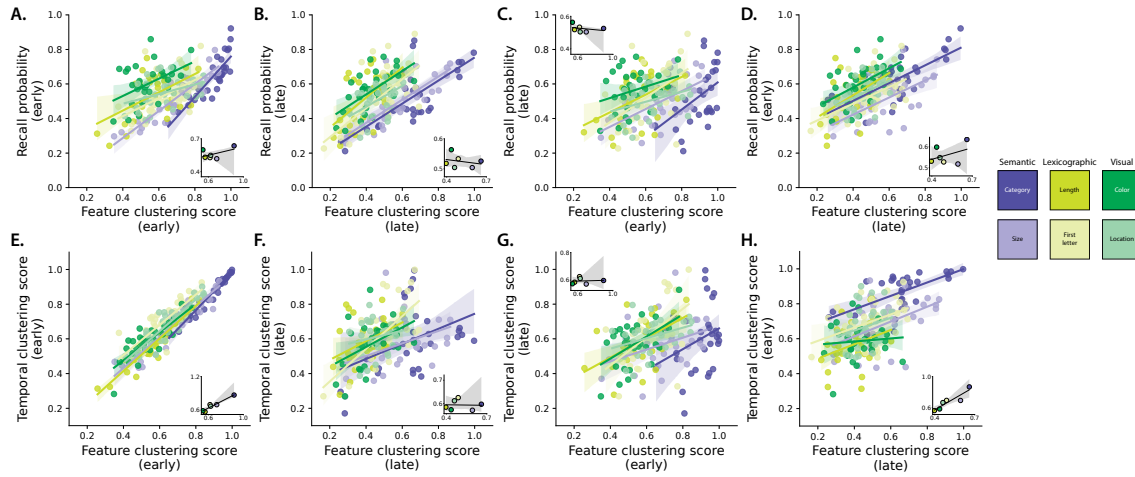


Figure 7: Interactions between feature clustering, recall probability, and contiguity. **A.** Recall probability versus feature clustering scores for order manipulation (early) lists. **B.** Recall probability versus feature clustering for randomly ordered (late) lists. **C.** Recall probability on late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

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

If participants show different sensitivities to order manipulations, how do their behaviors carry over to later lists? We found that participants who showed strong feature clustering on early lists often tended to show strong feature clustering on late lists (Fig. 8A; overall across participants and conditions: $r(179) = 0.592, p < 0.001$; non-visual feature

827 conditions: all $r_s \geq 0.350$, all $p_s \leq 0.058$; color: $r(29) = -0.071, p = 0.704$; location:
 828 $r(28) = 0.032, p = 0.868$; across conditions: $r(4) = 0.934, p = 0.006$). Although participants
 829 tended to show weaker feature clustering on late lists (Fig. 6) on *average*, the associations
 830 between early and late lists for individual participants suggests that some influence of
 831 early order manipulations may linger on late lists. We found that participants who exhib-
 832 ited larger carryover in feature clustering (i.e., continued to show strong feature clustering
 833 on late lists) for the semantic order manipulations (but not other manipulations) also
 834 tended to show a larger improvement in recall (Fig. 8B; overall: $r(179) = 0.378, p < 0.001$;
 835 category: $r(28) = 0.419, p = 0.021$; size: $r(28) = 0.737, p < 0.001$; non-semantic condi-
 836 tions: all $r_s \leq 0.252$, all $p_s \geq 0.179$; across conditions: $r(4) = 0.773, p = 0.072$) on late
 837 lists, relative to early lists. Participants who exhibited larger carryover in feature cluster-
 838 ing also tended to show stronger temporal clustering on late lists (relative to early lists)
 839 for all but the category condition (Fig. 8C; overall: $r(179) = 0.434, p < 0.001$; category:
 840 $r(28) = 0.229, p = 0.223$; all non-category conditions: all $r_s \geq 0.448$, all $p_s \leq 0.012$; across
 841 conditions: $r(4) = 0.598, p = 0.210$).

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

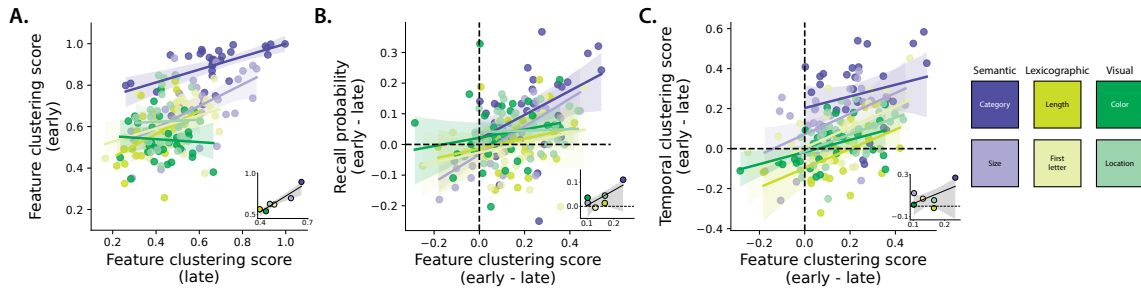


Figure 8: Feature clustering carryover effects. **A.** Feature clustering scores for order manipulation (early) versus randomly ordered (late) lists. **B.** Accuracy differences (on early versus late lists) versus feature clustering “carryover” (defined as the differences between the average clustering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

feature clustering from early to late lists. These participants might demonstrate strong recall performance not because of their inherently superior memory abilities, but rather because the specific condition they were assigned to happened to be especially easy for them, given their pre-experimental tendencies. To help distinguish between these interpretations, we designed an *adaptive* experimental condition (see *Adaptive condition*). The primary manipulation in the adaptive condition is that participants each experience three key types of lists. On *random* lists, words are ordered randomly (as in the feature rich condition). On *stabilize* lists, the presentation order is adjusted to be maximally similar to the current estimate of the participant’s memory fingerprint (see *Online “fingerprint” analysis*). Third, on *destabilize* lists, the presentation 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

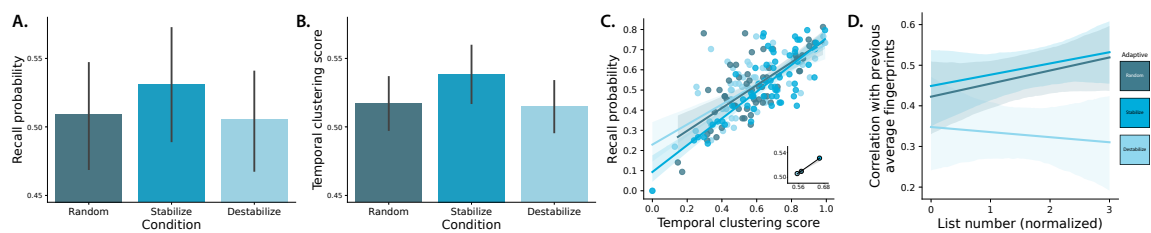


Figure 9: Adaptive free recall. **A.** Average probability of recall (taken across words, lists, and participants) for lists from each adaptive condition. **B.** Average temporal clustering scores for lists from each adaptive condition. **C.** Recall probability versus temporal clustering scores by participant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

are adjusted to best match each participant’s (potentially unique) memory fingerprint, the adaptive condition removes uncertainty about whether participants’ assigned conditions might just “happen” to match their preferred ways or organizing their memories.

Participants’ fingerprints on stabilize and random lists tended to become (numerically) slightly more similar to their average fingerprints computed from the previous lists they had experienced, and their fingerprints on destabilize lists tended to become numerically less similar (Fig. 9D). Overall, we found that participants tended to be better at remembering words on stabilize lists relative to words on random ($t(59) = 1.740, p = 0.087$) or destabilize ($t(59) = 1.714, p = 0.092$) lists (Fig. 9A). Participants showed no reliable differences in their memory performance on destabilize versus random lists ($t(59) = -0.249, p = 0.804$). Participants also exhibited stronger temporal clustering on stabilize lists, relative to random ($t(59) = 3.554, p = 0.001$) and destabilize ($t(59) = 4.045, p < 0.001$) lists (Fig. 9B). We found no reliable differences in temporal clustering for items on random versus destabilize lists ($t(59) = -0.781, p = 0.438$).

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 ($r(178) = 0.721, p < 0.001$), and for each list type individually (all $r_s \geq 0.683$, all $p_s \leq 0.001$). 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.

Discussion

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

905 randomly.

906 **Recap: visual feature manipulations**

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

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

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

928 **Recap: order manipulations**

929 When we (stochastically) sorted early lists along different feature dimensions, we found
930 several impacts on participants' memories. Sorting early lists semantically (by word cat-
931 egory) enhanced participants' memories for those lists, but the effects on performance of
932 sorting along other feature dimensions were inconclusive. However, each order manipu-
933 lation substantially affected how participants *organized* their memories of words from the
934 ordered lists. When we sorted lists semantically participants displayed stronger semantic
935 clustering; when we sorted lists lexicographically they displayed stronger lexicographic
936 clustering; and when we sorted lists visually they displayed stronger visual clustering.
937 Clustering along the unmanipulated feature dimensions in each of these cases was un-
938 changed.

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

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

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

953 in a random order). We also observed some (weaker) carryover effects of temporal cluster-
954 ing. Participants who showed stronger feature clustering (along their condition's feature
955 dimension) on early lists tended to show stronger temporal clustering on late lists. And
956 participants who showed stronger temporal clustering on early lists also tended to show
957 stronger feature clustering on late lists. Essentially, these order manipulations appeared
958 to affect each participant differently. Some participants were sensitive to our manipula-
959 tions, and those participants showed stronger impacts on their memory performance for
960 the ordered lists as well as future (random) lists. Other participants appeared relatively
961 insensitive to our manipulations, and those participants showed little carryover effects on
962 late lists.

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

978 encoding processes instead.

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

980 In real-world experience, each moment's unique blend of contextual features (where we
981 are, who we are with, what else we are thinking of at the time, what else we experience
982 nearby in time, etc.) plays an important role in how we interpret, experience, and re-
983 member that moment, and how we relate it to our other experiences (e.g., for review see
984 Manning, 2020). What are the analogues of real-world contexts in laboratory tasks like
985 the free recall paradigm employed in our study? In general, modern formal accounts of
986 free recall (Kahana, 2020) describe context as comprising a mix of (a) features pertaining
987 to or associated with each item and (b) other items and thoughts experienced nearby in
988 time, e.g., that might still be "lingering" in the participant's thoughts at the time they
989 study the item. Item features can include semantic properties (i.e., features related to the
990 item's meaning), lexicographic properties (i.e., features related to the item's letters), sen-
991 sory properties (i.e., feature related to the item's appearance, sound, smell, etc.), emotional
992 properties (i.e., features related to how meaningful the item is, whether the item evokes
993 positive or negative feelings, etc.), utility-related properties (e.g., features that describe
994 how an item might be used or incorporated into a particular task or situation), and more.
995 Essentially any aspect of the participant's experience that can be characterized, measured,
996 or otherwise described can be considered to influence the participant's mental context at
997 the moment they experience that item. Temporally proximal features include aspects of
998 the participant's internal or external experience that are *not* specifically occurring at the
999 moment they encounter an item, but that nonetheless influence how they process the item.
1000 Thoughts related to percepts, goals, expectations, other experiences, and so on that might
1001 have been cued (directly or indirectly) by the participant's recent experiences prior to the

1002 current moment all fall into this category. Internally driven mental states, such as thinking
1003 about an experience unrelated to the experiment, also fall into this category.

1004 Contextual features need not be intentionally or consciously perceived by the partic-
1005 ipant to affect memory, nor do they need to be relevant to the task instructions or the
1006 participant’s goals. Incidental factors such as font color (Jones and Pyc, 2014), background
1007 color (Isarida and Isarida, 2007), inter-stimulus images (Chiu et al., 2021; Gershman et al.,
1008 2013; Manning et al., 2016), background sounds (Beaman and Jones, 1998; Sahakyan and
1009 Smith, 2014), secondary tasks (Masicampo and Sahakyan, 2014; Polyn et al., 2009), and
1010 more can all impact how participants remember, and organize in memory, lists of studied
1011 items.

1012 Consistent with this prior work, we found that participants were sensitive to task-
1013 irrelevant visual features. We also found that changing the dynamics of those task-
1014 irrelevant visual features (in the reduced (early) and reduced (late) conditions) *also* affected
1015 participants’ memories. This suggests that it is not only the contextual features themselves
1016 that affect memory, but also the *dynamics* of context– i.e., how the contextual features
1017 associated with each item change over time.

1018 **Priming effects on memory performance and organization**

1019 When our ongoing experiences are ambiguous, we can draw on our past experiences,
1020 expectations, and other real, perceived, or inferred cues to help resolve the ambiguities.
1021 We may also be overtly or covertly “primed” to influence how we are likely to resolve
1022 ambiguities. For example, before listening to a story with several equally plausible inter-
1023 pretations, providing participants with “background” information beforehand can lead
1024 them towards one interpretation versus another (Yeshurun et al., 2017). More broadly, our
1025 conscious and unconscious biases and preferences can influence not only how we interpret

1026 high-level ambiguities, but even how we process low-level sensory information (Katabi
1027 et al., 2023).

1028 In more simplified scenarios, like list learning paradigms, the stimuli and tasks partic-
1029 ipants encounter before studying a given list can influence what and how they remember.
1030 For example, when participants are directed to suppress, disregard, or ignore “distracting”
1031 stimuli early on in an experiment, participants often tend to remember those stimuli less
1032 well when they are re-used as to-be-remembered targets later on in the experiment (Tip-
1033 per, 1985). In general, participants’ memories can be influenced by exposing them to
1034 a wide range of positive and negative priming factors before they encounter the to-be-
1035 remembered information (Balota et al., 1992; Clayton and Chattin, 1989; Donnelly, 1988;
1036 Flexser and Tulving, 1982; Gotts et al., 2012; Huang et al., 2004; Huber, 2008; Huber et al.,
1037 2001; McNamara, 1994; Neely, 1977; Rabinowitz, 1986; Tulving and Schacter, 1991; Watkins
1038 et al., 1992; Wiggs and Martin, 1998).

1039 The order manipulation conditions in our experiment show that participants can also be
1040 primed to pick up on more subtle statistical structure in their experiences, like the dynamics
1041 of how the presentation orders of stimuli vary along particular feature dimensions. These
1042 order manipulations affected not only how participants remembered the manipulated
1043 lists, but also how they remembered *future* lists with different (randomized) temporal
1044 properties.

1045 **Expectation, event boundaries, and situation models**

1046 Our findings that participants’ current and future memory behaviors are sensitive to
1047 manipulations in which features change over time, and how features change across items
1048 and lists, suggest parallels with studies on how we form expectations and predictions,
1049 segment our continuous experiences into discrete events, and make sense of different

scenarios and situations. Each of these real-world cognitive phenomena entail identifying statistical regularities in our experiences, and exploiting those regularities to gain insight, form inferences, organize or interpret memories, and so on. Our past experiences enable us to predict what is likely to happen in the future, given what happened “next” in our previous experiences that were similar to now (Barron et al., 2020; Brigard, 2012; Chow et al., 2016; Eichenbaum and Fortin, 2009; Gluck et al., 2002; Goldstein et al., 2021; Griffiths and Steyvers, 2003; Jones and Pashler, 2007; Kim et al., 2014; Manning, 2020; Tamir and Thornton, 2018; Xu et al., 2023).

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

1075 our experiences, and to generate expectations about how new experiences are likely to
1076 unfold. When those expectations are violated, we can perceive ourselves to have crossed
1077 into a new situation.

1078 In our study, we found that abruptly changing the “rules” about how the visual ap-
1079 pearances of words are determined, or about the orders in which words are presented,
1080 can lead participants to behave similarly to what one might expect upon crossing an event
1081 boundary. Adding in variability in font color and presentation locations for words on
1082 late lists, after those visual features had been held constant on early lists, led participants
1083 to remember more words on those later lists. One potential explanation is that partici-
1084 pants perceive an “event boundary” to have occurred when they encounter the first “late”
1085 list. According to contextual change accounts of memory across event boundaries (e.g., Sa-
1086 hakyan and Kelley, 2002), this could help to explain why participants in the reduced (early)
1087 and reduced (late) conditions exhibited better overall memory performance. Specifically,
1088 their memory for late list items could benefit from less interference from early list items,
1089 and the contextual features associated with late list items (after the “event boundary”)
1090 might serve as more specific recall cues for those late items (relative to if the boundary
1091 had not occurred).

1092 **Theoretical implications**

1093 Although most modern formal theories of episodic memory have been developed and
1094 tested to explain memory for list learning tasks (Kahana, 2020), a number of recent studies
1095 suggest some substantial differences between memory for lists versus naturalistic stim-
1096 uli (e.g., real-world experiences, narratives, films, etc.; Heusser et al., 2021; Lee et al., 2020;
1097 Manning, 2021; Nastase et al., 2020). One reason is that naturalistic stimuli are often much
1098 more engaging than the highly simplified list learning tasks typically employed in the

1099 psychological laboratory, perhaps leading participants to pay more attention, exert more
1100 effort, and stay more consistently motivated to perform well (Nastase et al., 2020). Another
1101 reason is that the temporal unfoldings of events and occurrences in naturalistic stimuli
1102 tend to be much more meaningful than the temporal unfoldings of items on typical lists
1103 used in laboratory memory tasks. Real-world events exhibit important associations at a
1104 broad range of timescales. For example, an early detail in a detective story may prove to
1105 be a clue to solving the mystery later on. Further, what happens in one moment typically
1106 carries some predictive information about what came before or after (Xu et al., 2023). In
1107 contrast, the lists used in laboratory memory tasks are most often ordered randomly, by
1108 design, to *remove* meaningful temporal structure in the stimulus (Kahana, 2012).

1109 On one hand, naturalistic stimuli provide a potential means of understanding how our
1110 memory systems function in the circumstances we most often encounter in our everyday
1111 lives. This implies that, to understand how memory works in the “real world,” we should
1112 study memory for stimuli that reflect the relevant statistical structure of real-world expe-
1113 riences. On the other hand, naturalistic stimuli can be difficult to precisely characterize or
1114 model, making it difficult to distinguish whether specific behavioral trends follow from
1115 fundamental workings of our memory systems, from some aspect of the stimulus, or from
1116 idiosyncratic interactions or interference between participants’ memory systems and the
1117 stimulus. This challenge implies that, to understand the fundamental nature of memory
1118 in its “pure” form, we should study memory for highly simplified stimuli that can pro-
1119 vide relatively unbiased (compared with real-world experiences) measures of the relevant
1120 patterns and tendencies.

1121 The experiment we report in this paper was designed to help bridge some of this gap
1122 between naturalistic tasks and more traditional list learning tasks. We had people study
1123 word lists similar to those used in classic memory studies, but we also systematically var-

1124 ied the lists' "richness" (by adding or removing visual features) and temporal structure
1125 (through order manipulations that varied over time and across experimental conditions).
1126 We found that participants' memory behaviors were sensitive to these manipulations.
1127 Some of the manipulations led to changes that were common across people (e.g., more
1128 temporal clustering when words' appearances were varied; enhanced memory for lists
1129 following an "event boundary;" more feature clustering on order-manipulated lists; etc.).
1130 Other manipulations led to changes that were idiosyncratic (especially carryover effects
1131 from order manipulations; e.g., participants who remembered more words on early order-
1132 manipulated lists tended to show stronger feature clustering for their condition's feature
1133 dimension on late randomly ordered lists; etc.). We also found that participants remem-
1134 bered more words from lists that were sorted to align with their idiosyncratic clustering
1135 preferences. Taken together, our results suggest that our memories are susceptible to ex-
1136 ternal influences (i.e., to the statistical structure of ongoing experiences), but the effects of
1137 past experiences on future memory are largely idiosyncratic across people.

1138 **Potential applications**

1139 Every participant in our study encountered exactly the same words, split into exactly the
1140 same lists. But participants' memory performance, the orders in which they recalled the
1141 words, and the effects of early list manipulations on later lists, varied according to how
1142 we presented the to-be-remembered words.

1143 Our findings raise a number of exciting questions. For example, how far might these
1144 manipulations be extended? In other words, might there be more sophisticated or clever
1145 feature or order manipulations that one might implement to have stronger impacts on
1146 memory? Are there limits to how much impact (on memory performance and/or or-
1147 ganization) these sorts of manipulations can have? Are those limits universal across

1148 people, or are there individual differences (based on prior experiences, natural strate-
1149 gies, neuroanatomy, etc.) that impose person-specific limits on the potential impact of
1150 presentation-level manipulations on memory?

1151 Our findings indicate that the ways word lists are presented affects how people re-
1152 member them. To the extent that word list memory reflects memory processes that are
1153 relevant to real-world experiences, one could imagine potential real-world applications of
1154 our findings. For example, we found that participants remembered more words when the
1155 presentation order agreed with their memory fingerprints. If analogous fingerprints could
1156 be estimated for classroom content, perhaps they could be utilized manually by teachers,
1157 or even by automated content presentation systems, to optimize how and what students
1158 remember.

1159 **Concluding remarks**

1160 Our work raises deep questions about the fundamental nature of human learning. What
1161 are the limits of our memory systems? How much does what we remember (and how we
1162 remember) depend on how we learn or experience the to-be-remembered content? We
1163 know that our expectations, strategies, situation models learned through prior experiences,
1164 and more, collectively shape how our experiences are remembered. But those aspects of
1165 our memory are not fixed: when we are exposed to the same experience in a new way, it
1166 can change how we remember that experience, and also how we remember, process, or
1167 perceive *future* experiences.

1168 **Author contributions**

1169 Conceptualization: JRM and ACH. Methodology: JRM and ACH. Software: JRM, KZ,
1170 PCF, CEF, and ACH. Analysis: JRM, PCF, and ACH. Data collection: KZ, EW, PCF, MRL,

1171 AMF, BJB, and CEF. Data curation and management: KZ, EW, PCF, MRL, ACH. Writing
1172 (original draft): JRM. Writing (review and editing): KZ, EW, PCF, MRL, AMF, BJB, CEF,
1173 and ACH. Supervision: JRM and ACH. Project administration: KZ, EW, and PCF. Funding
1174 acquisition: JRM.

1175 **Data and code availability**

1176 All of the data analyzed in this manuscript, along with all of the code for carrying out the
1177 analyses may be found at <https://github.com/ContextLab/FRFR-analyses>. Code for run-
1178 ning the non-adaptive experimental conditions may be found at [https://github.com/Con-](https://github.com/ContextLab/efficient-learning-code)
1179 [textLab/efficient-learning-code](https://github.com/ContextLab/efficient-learning-code). Code for running the adaptive experimental condition
1180 may be found at <https://github.com/ContextLab/adaptiveFR>. We have also released an as-
1181 sociated Python toolbox for analyzing free recall data, which may be found at [https://cdl-](https://cdl-quail.readthedocs.io/en/latest/)
1182 [quail.readthedocs.io/en/latest/](https://cdl-quail.readthedocs.io/en/latest/).

1183 **Acknowledgements**

1184 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) ver-
1185 sion of this study, and assistance with some of the data collection efforts from Rachel
1186 Chacko, Joseph Finkelstein, Sheherzad Mohyidin, Lucy Owen, Gal Perlman, Darya Ro-
1187 manova, Jake Rost, Jessica Tin, Marisol Tracy, and Peter Tran. Our work was supported
1188 in part by NSF CAREER Award Number 2145172 to JRM. The content is solely the re-
1189 sponsibility of the authors and does not necessarily represent the official views of our
1190 supporting organizations. The funders had no role in study design, data collection and
1191 analysis, decision to publish, or preparation of the manuscript.

References

- Anderson, J. R. and Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79(2):97–123.
- Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning and Motivation*, volume 2, pages 89–105. Academic Press, New York, NY.
- Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.
- Balota, D. A., Black, S. R., and Cheney, M. (1992). Automatic and attentional priming in young and older adults: reevaluation of the two-process model. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):485–502.
- Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a predictive coding account. *Progress in Neurobiology*, 192:101821–101834.
- Beaman, C. P. and Jones, D. M. (1998). Irrelevant sound disrupts order information in free recall as in serial recall. *The Quarterly Journal of Experimental Psychology Section A*, 51(3):615–636.
- Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 49:229–240.
- Bousfield, W. A., Sedgewick, C. H., and Cohen, B. H. (1954). Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology*, 67:111–118.
- Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11(2):177–220.

- 1214 Brigard, F. D. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*,
1215 3(420):1–3.
- 1216 Chiu, Y.-C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., and Sahakyan, L. (2021). Sepa-
1217 ration of item and context in item-method directed forgetting. *NeuroImage*, 235:117983.
- 1218 Chow, W.-Y., Momma, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory
1219 retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.
- 1220 Clayton, K. and Chaitin, D. (1989). Spatial and semantic priming effects in tests of spa-
1221 tial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
1222 15(3):495–506.
- 1223 Donnelly, R. E. (1988). Priming effects in successive episodic tests. *Journal of Experimental*
1224 *Psychology: Learning, Memory, and Cognition*, 14:256–265.
- 1225 DuBrow, S. and Davachi, L. (2016). Temporal binding within and across events. *Neurobi-*
1226 *ology of Learning and Memory*, 134:107–114.
- 1227 Eichenbaum, H. and Fortin, N. J. (2009). The neurobiology of memory based predictions.
1228 *Philosophical Transactions of the Royal Society of London Series B*, 364(1521):1183–1191.
- 1229 Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological*
1230 *Review*, 62:145–154.
- 1231 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory?
1232 *Psychological Science*, 22(2):243–252.
- 1233 Flexser, A. J. and Tulving, E. (1982). Priming and recognition failure. *Journal of Verbal*
1234 *Learning and Verbal Behavior*, 21:237–248.

- 1235 Gershman, S. J., Schapiro, A. C., Hupbach, A., and Norman, K. A. (2013). Neural context
1236 reinstatement predicts memory misattribution. *The Journal of Neuroscience*, 33(20):8590–
1237 8595.
- 1238 Glenberg, A. M., Bradley, M. M., Kraus, T. A., and Renzaglia, G. J. (1983). Studies of the
1239 long-term recency effect: support for a contextually guided retrieval theory. *Journal of*
1240 *Experimental Psychology: Learning, Memory, and Cognition*, 12:413–418.
- 1241 Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
1242 prediction” task? individual variability in strategies for probabilistic category learning.
1243 *Learning and Memory*, 9:408–418.
- 1244 Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder,
1245 A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto,
1246 C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A.,
1247 Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. (2021). Thinking
1248 ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*,
1249 page doi.org/10.1101/2020.12.02.403477.
- 1250 Gotts, S. J., Chow, C. C., and Martin, A. (2012). Repetition priming and repetition sup-
1251 pression: A case for enhanced efficiency through neural synchronization. *Cognitive*
1252 *Neuroscience*, 3(3-4):227–237.
- 1253 Griffiths, T. L. and Steyvers, M. (2003). Prediction and semantic association. *Advances in*
1254 *Neural Information Processing Systems*, 15.
- 1255 Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml,
1256 M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages
1257 2338–2342.

- 1258 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail:
1259 a Python toolbox for analyzing and plotting free recall data. *Journal of Open Source*
1260 *Software*, 10.21105/joss.00424.
- 1261 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal
1262 behavioral and neural signatures of transforming naturalistic experiences into episodic
1263 memories. *Nature Human Behavior*, 5:905–919.
- 1264 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). HyperTools: a
1265 Python toolbox for gaining geometric insights into high-dimensional data. *Journal of*
1266 *Machine Learning Research*, 18(152):1–6.
- 1267 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context.
1268 *Journal of Mathematical Psychology*, 46:269–299.
- 1269 Huang, L., Holcombe, A. O., and Pashler, H. (2004). Repetition priming in visual search:
1270 episodic retrieval, not feature priming. *Memory and Cognition*, 32:12–20.
- 1271 Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental*
1272 *Psychology: General*, 137(2):324–347.
- 1273 Huber, D. E., Shiffrin, R. M., Lyle, K. B., and Ruys, K. I. (2001). Perception and preference
1274 in short-term word priming. *Psychological Review*, 108(1):149–182.
- 1275 Isarida, T. and Isarida, T. K. (2007). Environmental context effects of background color in
1276 free recall. *Memory and Cognition*, 35(7):1620–1629.
- 1277 Jenkins, J. J. and Russell, W. A. (1952). Associative clustering during recall. *Journal of*
1278 *Abnormal and Social Psychology*, 47:818–821.

- 1279 Jones, A. C. and Pyc, M. A. (2014). The production effect: costs and benefits in free recall.
1280 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):300–305.
- 1281 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing
1282 prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 1283 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*,
1284 24:103–109.
- 1285 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York,
1286 NY.
- 1287 Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychol-*
1288 *ogy*, 71:107–138.
- 1289 Kahana, M. J., Howard, M. W., and Polyn, S. M. (2008). Associative processes in episodic
1290 memory. In Roediger III, H. L., editor, *Cognitive Psychology of Memory*, pages 476–490.
1291 Elsevier, Oxford, UK.
- 1292 Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., and Yeshurun, Y. (2023). Deeper than
1293 you think: partisanship-dependent brain responses in early sensory and motor brain
1294 regions. *The Journal of Neuroscience*, pages doi.org/10.1523/JNEUROSCI.0895–22.2022.
- 1295 Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning
1296 of memories by context-based prediction error. *Proceedings of the National Academy of*
1297 *Sciences, USA*, In press.
- 1298 Kimball, D. R., Smith, T. A., and Kahana, M. J. (2007). The fSAM model of false recall.
1299 *Psychological Review*, 114(4):954–993.

- 1300 Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases
1301 of human memory. *Current Opinion in Behavioral Sciences*, 32:111–119.
- 1302 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
1303 *Handbook of Human Memory*. Oxford University Press.
- 1304 Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
1305 function? *Psychological Review*, 128(4):711–725.
- 1306 Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. A.
1307 (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic*
1308 *Bulletin and Review*, 23(5):1534–1542.
- 1309 Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free
1310 recall. *Memory*, 20(5):511–517.
- 1311 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic
1312 memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.
- 1313 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory pat-
1314 terns in temporal lobe reveal context reinstatement during memory search. *Proceedings*
1315 *of the National Academy of Sciences, USA*, 108(31):12893–12897.
- 1316 Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., and Kahana, M. J. (2012).
1317 Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clus-
1318 tering during memory search. *The Journal of Neuroscience*, 32(26):8871–8878.
- 1319 Masicampo, E. J. and Sahakyan, L. (2014). Imagining another context during encoding off-
1320 sets context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory,*
1321 *and Cognition*, 40(6):1772–1777.

- 1322 Masís-Obando, R., Norman, K. A., and Baldassano, C. (2022). Scheme representations in
 1323 distinct brain networks support narrative memory during encoding and retrieval. *eLife*,
 1324 11:e70445.
- 1325 McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental*
 1326 *Psychology: Learning, Memory, and Cognition*, 20:507–520.
- 1327 Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman,
 1328 S. J. (2017). The successor representation in human reinforcement learning. *Nature*
 1329 *Human Behavior*, 1:680–692.
- 1330 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental*
 1331 *Psychology: General*, 64:482–488.
- 1332 Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy
 1333 of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1334 Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhi-
 1335 bitionless spreading activation and limited-capacity attention. *Journal of Experimental*
 1336 *Psychology: General*, 106(3):226–254.
- 1337 Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of
 1338 context. *Trends in Cognitive Sciences*, 12:24–30.
- 1339 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). Task context and organization in
 1340 free recall. *Neuropsychologia*, 47:2158–2163.
- 1341 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly*
 1342 *Journal of Experimental Psychology*, 17:132–138.

- 1343 Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of
1344 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,
1345 NY.
1346
- 1347 Rabinowitz, J. C. (1986). Priming in episodic memory. *Journal of Gerontology*, 41:204–213.
- 1348 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:
1349 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.
- 1350 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition.
1351 *Current Opinion in Behavioral Sciences*, 17:133–140.
- 1352 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior.
1353 *Nature Reviews Neuroscience*, 13:713–726.
- 1354 Romney, A. K., Brewer, D. D., and Batchelder, W. H. (1993). Predicting clustering from
1355 semantic structure. *Psychological Science*, 4:28–34.
- 1356 Sahakyan, L. and Kelley, C. M. (2002). A contextual change account of the directed
1357 forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
1358 28(6):1064–1072.
- 1359 Sahakyan, L. and Smith, J. R. (2014). A long time ago, in a context far, far away: Retro-
1360 spective time estimates and internal context change. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):86–93.
1361
- 1362 Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclo-*
1363 *pedic Reference*, 3:501–506.

- 1364 Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008). A context-based theory of
1365 recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.
- 1366 Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of
1367 time. *Neural Computation*, 24:134–193.
- 1368 Sirotin, Y. B., Kimball, D. R., and Kahana, M. J. (2005). Going beyond a single list: modeling
1369 the effects of prior experience on episodic free recall. *Psychonomic Bulletin and Review*,
1370 12(5):787–805.
- 1371 Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: a review and
1372 meta-analysis. *Psychonomic Bulletin and Review*, 8(2):203–220.
- 1373 Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., and Zacks, J. M. (2011).
1374 Changes in events alter how people remember recent information. *Journal of Cognitive*
1375 *Neuroscience*, 23(5):1052–1064.
- 1376 Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception
1377 affect memory encoding and updating. *Journal of Experimental Psychology: General*,
1378 138(2):236–257.
- 1379 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in*
1380 *Cognitive Sciences*, 22(3):201–212.
- 1381 Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *The*
1382 *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37:571–
1383 590.
- 1384 Tulving, E. and Schacter, D. L. (1991). Priming and human memory systems. *Science*,
1385 247:301–305.

- 1386 Watkins, P. C., Mathews, A., Williamson, D. A., and Fuller, R. D. (1992). Mood-congruent
1387 memory in depression: emotional priming or elaboration? *Journal of Abnormal Psychol-*
1388 *ogy*, 101(3):581–586.
- 1389 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American*
1390 *Journal of Psychology*, 35:396–401.
- 1391 Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming.
1392 *Current Opinion in Neurobiology*, 8(2):227–233.
- 1393 Xu, X., Zhu, Z., and Manning, J. R. (2023). The psychological arrow of time drives
1394 temporal asymmetries in retrodicting versus predicting narrative events. *PsyArXiv*,
1395 page doi.org/10.31234/osf.io/yp2qu.
- 1396 Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., and Hasson, U.
1397 (2017). Same story, different story: the neural representation of interpretive frameworks.
1398 *Psychological Science*, 28(3):307–319.
- 1399 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018).
1400 Is automatic speech-to-text transcription ready for use in psychological experiments?
1401 *Behavior Research Methods*, 50:2597–2605.
- 1402 Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation
1403 models in narrative comprehension: an event-indexing model. *Psychological Science*,
1404 6(5):292–297.
- 1405 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension
1406 and memory. *Psychological Bulletin*, 123(2):162–185.