

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

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

¹Dartmouth College

²Akili Interactive Labs

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

Abstract

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

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

Author contributions. Conceptualization: JRM and ACH. Data curation: JRM, PCF, ACH. Formal analysis: JRM, PCF, and ACH. Funding acquisition: JRM. Investigation: ECW, PCF, MRL, AMF, BJB, DR, CEF, and ACH. Methodology: JRM and ACH. Project administration: ECW and PCF. Resources: JRM. Software: JRM, PCF, CEF, and ACH. Supervision: JRM and ACH. Validation: JRM, PCF, and ACH. Writing (original draft): JRM. Writing (review and editing): ECW, PCF, MRL, AMF, BJB, DR, CEF, and ACH.

Author note. All of the data analyzed in this manuscript, along with all of the code for carrying out the analyses may be found at <https://github.com/ContextLab/FRFR-analyses>. Note that this study was not preregistered. Some of the ideas and data presented in this manuscript were also presented at the Annual Meeting of the Society for Neuroscience (2017) and the Context and Episodic Memory Symposium (2017).

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

35 Introduction

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

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

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

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

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

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

85 is an example of temporal clustering, which is perhaps the dominant form of organization
86 in free recall. This effect can be seen in people’s tendencies to successively recall items that
87 occupied neighboring positions in the studied list (Kahana, 1996). There are also striking
88 effects of semantic clustering (Bousfield, 1953; Bousfield et al., 1954; Jenkins and Russell,
89 1952; Manning and Kahana, 2012; Romney et al., 1993), whereby the recall of a given
90 item is more likely to be followed by recall of a similar or related item than a dissimilar
91 or unrelated one. In general, people organize memories for words along a wide variety
92 of stimulus dimensions. According to models like the *Context Maintenance and Retrieval*
93 model (Polyn et al., 2009), the stimulus features associated with each word (e.g., the word’s
94 meaning, size of the object the word represents, letters that make up the word, font size,
95 font color, location on the screen, etc.) are incorporated into the participant’s mental con-
96 text representation (Manning, 2020; Manning et al., 2015, 2011, 2012; Smith and Vela, 2001).
97 During a memory test, any of these features may serve as a memory cue, which in turn
98 leads the participant to successively recall words that share stimulus features.

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

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

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

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

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

145 **Materials and methods**

146 **Participants**

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

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

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

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

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

183 five participants declined to report their race. We note that several participants reported
184 more than one of the above racial categories. Participants reported their highest degrees
185 achieved as “Some college” (359 participants), “High school graduate” (117 participants),
186 “College graduate” (seven participants), “Some high school” (five participants), “Doctor-
187 ate” (one participant), and “Master’s degree” (one participant). A total of 482 participants
188 reported no reading impairments; eight reported having mild reading impairments. A
189 total of 489 participants reported having normal color vision and one participant reported
190 having impaired color vision. A total of 482 participants reported taking no prescrip-
191 tion medications and having no recent injuries; four participants reported having ADHD,
192 one reported having dyslexia, one reported having allergies, one reported a recently torn
193 ACL/MCL, and one reported a concussion from several months prior. The participants
194 reported having consumed 0–3 cups of coffee on the day of the testing session (mean: 0.32
195 cups; standard deviation: 0.58 cups). Participants reported their current level of alertness,
196 and we converted their responses to numerical scores as follows: “very sluggish” (-2),
197 “a little sluggish” (-1), “neutral” (0), “a little alert” (1), and “very alert” (2). Across all
198 participants, the full range of alertness levels were reported (range: -2–2; mean: 0.35;
199 standard deviation: 0.89).

200 We dropped from our dataset the one participant who reported having abnormal color
201 vision, as well as 38 participants whose data were corrupted due to technical failures while
202 running the experiment or during the daily database merges. In total, this left usable data
203 from 452 participants, broken down by experimental condition as follows: feature-rich (67
204 participants), reduced (61 participants), reduced (early) (42 participants), reduced (late)
205 (41 participants), category (30 participants), size (30 participants), length (30 participants),
206 first letter (30 participants), color (31 participants), location (30 participants), and adaptive
207 (60 participants). The participant who declined to fill out their demographic survey

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

210 **Experimental design**

211 Our experiment is a variant of the classic free recall paradigm that we term “*feature-*
212 *rich free recall*.” In feature-rich free recall, participants study 16 lists, each comprised
213 of 16 words that vary along a number of stimulus dimensions (Fig. 1). The stimulus
214 dimensions include two semantic features related to the *meanings* of the words (semantic
215 category, referent object size), two lexicographic features related to the *letters* that make
216 up the words (word length in number of letters, identity of the word’s first letter), and
217 two visual features that are independent of the words themselves (font color, presentation
218 location). Each list contains four words from each of four different semantic categories,
219 with two referent object sizes reflected across all of the words. After studying each
220 list, the participant attempts to recall as many words as they can from that list, in any
221 order they choose. Because each individual word is associated with several well-defined
222 (and quantifiable) features, and because each list incorporates a diverse mix of feature
223 values along each dimension, this allows us to estimate which features participants are
224 considering or leveraging in organizing their memories. Our study was approved by
225 the Dartmouth College Committee for the Protection of Human Subjects (CPHS) under a
226 protocol titled “*Efficient Learning*.”

227 **Stimuli**

228 The stimuli in our paradigm were 256 English words selected in a previous study (Ziman
229 et al., 2018). All words referred to concrete nouns and were chosen from 15 unique semantic
230 categories: body parts, building-related, cities, clothing, countries, flowers, fruits, insects,



Figure 1: Feature-rich free recall. After studying lists comprised of words that vary along several feature dimensions, participants verbally recall words in any order (microphone icon). Each experimental condition manipulates word features and/or presentation orders within and/or across lists. The rows display representative (illustrated) examples of items from the first list participants might encounter in each condition. The rectangles during the “Presentation phase” show illustrated screen captures during a series of word presentations. Each word appeared onscreen for 2 seconds, followed by 2 seconds of blank screen. The red microphone icons during the “Recall” phase denote the one-minute verbal recall interval. The labels on the right (and corresponding groupings on the left) denote experimental condition labels.

231 instruments, kitchen-related, mammals, (US) states, tools, trees, and vegetables. We also
232 tagged each word according to the approximate size of the object it referred to. Words
233 were labeled as “small” if the referent object was likely able to “fit in a standard shoebox”
234 or “large” if the object was larger than a shoebox. Most semantic categories comprised
235 words that reflected both “small” and “large” object sizes, but several included only one
236 or the other (e.g., all countries, US states, and cities are larger than a shoebox; mean
237 number of different sizes per category: 1.33; standard deviation: 0.49). The number of
238 words in each semantic category also varied from 12–28 (mean number of words per
239 category: 17.07; standard deviation: 4.65). We also identified lexicographic features for
240 each word, including its first letter and length (i.e., number of letters). Across all categories,
241 all possible first letters were represented except for ‘Q’ (average number of unique first
242 letters per category: 11.00; standard deviation: 2.00 letters). Word lengths ranged from
243 3–12 letters (average: 6.17 letters; standard deviation: 2.06 letters).

244 We assigned the categorized words into a total of 16 lists with several constraints. First,
245 we required that each list contain exactly four unique words from each of four unique
246 categories. Second, we required that each list contain at least one word representing each
247 of the two object sizes (“small” and “large”). On average, each category was represented in
248 4.27 lists (standard deviation: 1.16 lists). Aside from these two constraints, we randomly
249 assigned each word to a single list (i.e., such that no words appeared in multiple lists
250 or were omitted entirely). After random assignment, each list contained words with an
251 average of 11.13 unique starting letters (standard deviation: 1.15 letters) and an average
252 length of 6.17 letters (standard deviation: 0.34 letters).

253 The above assignments of words to lists was performed once across all participants,
254 such that every participant studied the same set of 16 lists. In every condition, we
255 randomized the study order of these lists across participants. For participants in most

256 conditions, on some or all of the lists, we also randomly varied two additional visual
257 features associated with each word: the presentation font color and the word’s onscreen
258 location. These attributes were assigned independently for each word (and for every
259 participant). These visual features were varied for words in all lists and conditions except
260 for the “reduced” condition (all lists), the first eight lists of the “reduced (early)” condition,
261 and the last eight lists of the “reduced (late)” condition. In these latter cases, all words
262 were presented in black at the center of the experimental computer’s display.

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

271 Most of the experimental manipulations we carried out entailed presenting or sorting
272 the presented words differently on the first eight lists participants studied (which we call
273 “*early*” lists) versus on the final eight lists they studied (“*late*” lists). Since every participant
274 studied exactly 16 lists, every list was either “*early*” or “*late*” depending on its order in
275 the list study sequence. (In other words, the “*early*” and “*late*” labels capture all of the
276 lists participants studied.)

277 **Real-time speech-to-text processing**

278 Our experimental paradigm incorporates the Google Cloud Speech-to-Text engine (Halpern
279 et al., 2016) to automatically transcribe participants’ verbal recalls into text. This allows

280 recalls to be transcribed in real time—a distinguishing feature of the experiment; in typical
281 verbal recall experiments, the audio data must be parsed and transcribed manually. In
282 prior work, we used a similar experimental setup (equivalent to the “reduced” condition
283 in the present study) to verify that the automatically transcribed recalls were sufficiently
284 close to human-transcribed recalls to yield reliable data (Ziman et al., 2018). This real-time
285 speech processing component of the paradigm plays an important role in the “adaptive”
286 condition of the experiment, as described below.

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

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

296 We also designed two conditions in which we varied the words’ visual appearances
297 across lists. In the *reduced (early)* condition, we followed the “reduced” procedure (pre-
298 senting each word in black at the center of the screen) for early lists, and followed the
299 “feature-rich” procedure (presenting each word in a random color and location) for late
300 lists. Finally, in the *reduced (late)* condition, we followed the feature-rich procedure for
301 early lists and the reduced procedure for late lists.

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

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

312 **Defining feature-based distances.** Sorting words according to a given relevant feature
313 requires first defining a distance function for quantifying the dissimilarity between the
314 values of that feature for each pair of words. This function varied according to the type
315 of feature under consideration. Semantic features (category and size) are *categorical*. For
316 these features, we defined a binary distance function: two words were considered to
317 “match” (i.e., have a distance of 0) if their labels were the same (i.e., both from the same
318 semantic category or both of the same size). If two words' labels were different for a given
319 feature, we defined the words to have a distance of 1. Lexicographic features (length
320 and first letter) are *discrete*. For these features, we defined a discrete distance function.
321 Specifically, we defined the distance between two words as either the absolute difference
322 between their lengths, or the absolute distance between their starting letters in the English
323 alphabet, respectively. For example, two words that started with the same letter would
324 have a “first letter” distance of 0, and a pair of words starting with ‘J’ and ‘A’ would have
325 a first letter distance of 9. Because words' lengths and letters' positions in the alphabet

are always integers, these discrete distances always take on integer values. Finally, the visual features (color and location) are *continuous* and *multivariate*, in that each “feature” is defined by multiple (positive) real values. We defined the “color” and “location” distances between two words as the Euclidean distances between their (r, g, b) color vectors and (x, y) location vectors (specified as a percentage of the viewable display’s width), respectively. Therefore, the color and location distance measures always take on non-negative real values (upper-bounded at 439.94 for color, or 124.52 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 words on the to-be-presented list. Second, we compute the distances between the chosen word’s feature value(s) and the corresponding feature value(s) of all yet-to-be-presented words. Third, we convert these distances (between the previously presented word’s feature values, a , and each of the W yet-to-be-presented candidate words’ feature values, $b_{i \in 1 \dots W}$) to similarity scores:

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

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

361 Adaptive condition

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

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

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

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

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

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

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

427 Following our prior work (Heusser et al., 2017), we used a permutation-based cor-
428 rection procedure to help isolate the behavioral aspects of clustering that we were most
429 interested in. After computing the uncorrected clustering score (for the given list and
430 observed recall sequence), we constructed a “null” distribution of n additional clustering
431 scores by repeatedly randomly shuffling the order of the recalled words and recomputing
432 the clustering score for these shuffled recall sequences (we use $n = 500$ in the present

study). This null distribution represents an approximation of the range of clustering scores one might expect to observe by “chance,” given that a hypothetical participant was *not* truly clustering their recalls, but where the hypothetical participant still studied and recalled exactly the same items (with the same features) as the true participant. We define the *permutation-corrected clustering score* as the percentile rank of the observed uncorrected clustering score in this estimated null distribution. In this way, a corrected score of 1 indicates that the observed score was greater than any clustering score one might expect by chance—in other words, good evidence that the participant was truly clustering their recalls along the given feature dimension. We applied this correction procedure to all of the clustering scores (feature and temporal) reported in this paper. In Figure S4, we report how participants’ clustering scores along different feature dimensions (in the order manipulation conditions) are correlated, and how clustering scores change across lists.

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

Online “fingerprint” analysis. The presentation orders of some lists in the adaptive condition of our experiment (see *Adaptive condition*) were sorted according to each individual participant’s *current* memory fingerprint, estimated using all of the lists they had studied

up to that point in the experiment. Because our experiment incorporated a speech-to-text component, all of the behavioral data for each participant could be analyzed just a few seconds after the conclusion of the recall intervals for each list. We used the Quail Python package (Heusser et al., 2017) to apply speech-to-text algorithms to the just-collected audio data, aggregate the data for the given participant, and estimate the participant’s memory fingerprint using all of their available data up to that point in the experiment. Two aspects of our implementation are worth noting: First, because memory fingerprints are computed independently for each list and then averaged across lists, the already-computed memory fingerprints for earlier lists could be cached and retrieved as needed in future computations. This meant that updating our estimate of a participant’s memory fingerprint required computing feature and temporal clustering scores only for the single most recent list. Second, the clustering scores for each dimension of a participant’s memory fingerprint could be estimated independently from the others, as could each element of the null distributions of uncorrected clustering scores computed for each dimension (see *Permutation-corrected feature clustering scores*). This enabled us to aggressively parallelize the fingerprint-updating procedure and compress the relevant computations into just a few seconds of computing time. The combined processing time for the speech-to-text algorithm, fingerprint computations, and permutation-based ordering procedure (described next) easily fit within the inter-list intervals, where participants paused for a self-paced break before moving on to study and recall the next list.

Ordering “stabilize” and “destabilize” lists by an estimated fingerprint. In the adaptive condition of our experiment, the presentation orders for *stabilize* and *destabilize* lists were chosen to either maximally or minimally (respectively) comport with participants’ memory fingerprints. Given a participant’s memory fingerprint and a to-be-presented set of items, we designed a permutation-based procedure for ordering the items. First, we

dropped from the participant’s fingerprint the temporal clustering score. For the remaining feature dimensions, we arranged the clustering scores in the fingerprint into a template vector f . Second, we computed $n = 2500$ random permutations of the to-be-presented items. These permutations served as candidate presentation orders. We sought to select the specific order that most (or least) closely matched f . Third, for each random permutation, we computed the (permutation-corrected) “fingerprint,” treating the permutation as though it were a potential “perfect” recall sequence. (We did not include temporal clustering scores in these fingerprints, since the temporal clustering score for every permutation is always equal to 1.) This yielded a “simulated fingerprint” vector \hat{f}_p for each permutation p . We used these simulated fingerprints to select a specific permutation i that either maximized (for stabilize lists) or minimized (for destabilize lists) the correlation between \hat{f}_i and f .

Computing low-dimensional embeddings of memory fingerprints

Following some of our prior work (Fitzpatrick et al., 2023; Heusser et al., 2021, 2018; Manning et al., 2022), we used low-dimensional embeddings to help visualize how participants’ memory fingerprints change across lists (Figs. 6A, S8A). To compute a shared embedding space across participants and experimental conditions, we concatenated the full set of across-participant average fingerprints (for all lists and experimental conditions) to create a large matrix with number-of-lists (16) \times number-of-conditions (10 , excluding the adaptive condition) rows and seven columns (one for each feature clustering score, plus an additional temporal clustering score column). We used principal components analysis to project the seven-dimensional observations into a two-dimensional space (using the two principal components that explained the most variance in the data). For two visualizations (Figs. 6B, S8B), we computed an additional set of two-dimensional embeddings for the

506 *average* fingerprints across lists within a given list grouping (i.e., early or late). For those
507 visualizations, we averaged across the rows (for each condition and group of lists) in the
508 combined fingerprint matrix prior to projecting it into the shared two-dimensional space.
509 This yielded a single two-dimensional coordinate for each *list group* (in each condition),
510 rather than for each individual list. We used these embeddings solely for visualization;
511 all statistical tests were carried out in the original (seven-dimensional) feature spaces.

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

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

521 We developed a permutation-based procedure to factor out the effects of temporal
522 clustering from the clustering scores for each feature dimension. For a given set of recalled
523 items (whose presentation positions are given by $x_1, x_2, x_3, \dots, x_L$), we circularly shifted
524 the presentation positions by a randomly chosen amount (between 1 and the list length,
525 L) to obtain a new set of items at the (now altered) positions of the original recalls.
526 Since the new set of items will have the same (average) temporal distances between
527 successive recalls, the temporal clustering score for the new set of items will be equal (on
528 average) to the temporal clustering score for the original recalls. However, we can then
529 re-compute the feature clustering score for those new items. Finally, we can compute a

530 “temporally corrected” feature clustering score by computing the average percentile rank
531 of the observed (raw) feature clustering score within the distributions of circularly shifted
532 feature clustering scores, across $n = 500$ repetitions of this procedure. This new temporally
533 corrected score provides an estimate of the observed degree of feature clustering over and
534 above what could be accounted for by temporal clustering alone.

535 While these temporally corrected clustering scores are useful for identifying when
536 feature clustering cannot be accounted for by temporal clustering alone, they are *not*
537 necessarily valid estimates of the “true” degree to which participants are organizing their
538 memories along a given feature dimension. For example, on a list where the presentation
539 order and feature values (along the given feature dimension) are perfectly correlated, the
540 temporally corrected score will have an expected value of 0.5 no matter which words a
541 participant recalls, or the order in which they recall them. Therefore these temporally
542 corrected clustering scores are interpretable only to the extent that presentation order and
543 feature value are decoupled.

544 **Analyses**

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

546 Probability of first recall curves (Atkinson and Shiffrin, 1968; Postman and Phillips, 1965;
547 Welch and Burnett, 1924) reflect the probability that an item will be recalled first, as
548 a function of its serial position during encoding. We used an analogous approach to
549 compute the proportion of trials on which each item (as a function of its presentation
550 position) was recalled at each output position n (Hogan, 1975; Howard and Kahana, 1999;
551 Polyn et al., 2009; Zhang et al., 2023). To carry out this analysis, we initialized (for each
552 participant) a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then,
553 for each list, we found the presentation index of the word that was recalled first, and filled

554 in that position in the matrix with a 1. Finally, we averaged over the rows of the matrix
555 to obtain a 1 by 16 array of probabilities, for each participant. We used an analogous
556 procedure to compute probability of n^{th} recall curves for each participant. Specifically,
557 we filled in the corresponding matrices according to the n^{th} recall on each list that each
558 participant made. When a given participant had made fewer than n recalls for a given
559 list, we simply excluded that list from our analysis when computing that participant's
560 curve(s). The probability of first recall curve corresponds to a special case where $n = 1$.

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

570 **Lag-conditional response probability curve**

571 The lag-conditional response probability (lag-CRP) curve (Kahana, 1996) reflects the prob-
572 ability of recalling a given item after the just-recalled item, as a function of the items'
573 relative encoding positions (lag). In other words, a lag of 1 indicates that a recalled item
574 was presented immediately after the previously recalled item, and a lag of -3 indicates
575 that a recalled item came three items before the previously recalled item. For each re-
576 call transition (following the first recall), we computed the lag between the presentation
577 positions of the just-recalled word and the next-recalled word. We then computed the

578 proportions of transitions (between successively recalled words) for each lag, normalizing
579 for the total numbers of possible transitions. In carrying out this analysis, we excluded
580 all incorrect recalls and repetitions (i.e., recalling a word that had already appeared in
581 the current recall sequence). This yielded, for each list, a 1 by number-of-lags (–15 to
582 +15; 30 lags in total, excluding lags of 0) array of conditional probabilities. We averaged
583 these probabilities across lists to obtain a single lag-CRP for each participant. Because
584 transitions at large absolute lags are rare, these curves are typically displayed using range
585 restrictions (Kahana, 2012).

586 **Serial position curve**

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

597 **Identifying event boundaries**

598 We used the distances between feature values for successively presented words (see *Defin-*
599 *ing feature-based distances*) to estimate “event boundaries” where the feature values changed
600 more than usual (DuBrow and Davachi, 2016; Ezzyat and Davachi, 2011; Manning et al.,

2016; Radvansky and Copeland, 2006; Swallow et al., 2011, 2009). For each list, for each feature dimension, we computed the distribution of distances between the feature values for successively presented words. We defined event boundaries (e.g., Fig. 3B) as occurring between any successive pair of words whose distances along the given feature dimension were greater than two standard deviations above the mean for that list. Note that, because event boundaries are defined for each feature dimension, each individual list may contain several sets of event boundaries, each at different moments in the presentation sequence (depending on the feature dimension of interest).

Transparency and openness

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

Results

While holding the set of words (and the assignments of words to lists) constant, we manipulated two aspects of participants' experiences of studying each list. We sought to understand the effects of these manipulations on participants' memories for the studied words. First, we added two additional sources of visual variation to the individual word presentations: font color and onscreen location. Importantly, these visual features were independent of the meaning or semantic content of the words (e.g., word category, size

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

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

of the referent, etc.) and of the lexicographic properties of the words (e.g., word length, first letter, etc.). We wondered whether this additional word-independent information might facilitate recall (e.g., by providing new or richer potential ways of organizing or retrieving memories of the studied words; Davachi et al., 2003; Drewnowski and Murdock, 1980; Hargreaves et al., 2012; Madan, 2021; Meinhardt et al., 2020; Slamecka and Barlow, 1979; Socher et al., 2009) or impair recall (e.g., by distracting or confusing participants with irrelevant information; Lange, 2005; Marsh et al., 2012, 2015; Reinitz et al., 1992). Second, we manipulated the orders in which words were studied (and how those orderings changed over time). We wondered whether presenting the same list of words with different appearances (e.g., by manipulating font size and onscreen location) or in different orders (e.g., sorted along one feature dimension versus another) might serve to influence how participants organized their memories of the words (e.g., Manning et al., 2015; Polyn and Kahana, 2008). We also wondered whether some order manipulations might be temporally “sticky” by influencing how *future* lists were remembered (e.g., Baddeley, 1968; Darley and Murdock, 1971; Lohnas et al., 2010; Sirotin et al., 2005; Whitely, 1927).

To obtain a clean preliminary estimate of the consequences on memory of randomly varying the font colors and locations of presented words (versus holding the font color fixed at black, and the words’ locations fixed at the center of the screen), we compared participants’ performance on the *feature-rich* and *reduced* experimental conditions (see *Random*

643 conditions, Fig. S1, Tab. 1). In the feature-rich condition, the words' colors and locations var-
644 ied randomly, and in the reduced condition, words were always presented in black, at the
645 center of the display. Aggregating across all lists for each participant, we found no differ-
646 ence in recall accuracy (i.e., the proportions of words successfully recalled) for feature-rich
647 versus reduced lists. However, participants in the feature-rich condition clustered their
648 recalls substantially more along every dimension we examined (see *Permutation-corrected*
649 *feature clustering scores* for more information about how we quantified each participant's
650 clustering tendencies.) Taken together, these comparisons suggest that adding new fea-
651 tures changes how participants organize their memories of studied words, even when
652 those new features are independent of the words themselves and vary randomly across
653 words. We found no evidence that those additional uninformative features were distract-
654 ing (in terms of their impact on memory performance), but they did affect participants'
655 recall dynamics (measured via their clustering scores).

656 A core assumption of our approach is that each participant organizes their memo-
657 ries in a unique way. We defined each participant's *memory fingerprint* as the set of their
658 permutation-corrected clustering scores across all dimensions we tracked in our study,
659 including their six feature-based clustering scores (category, size, length, first letter, color,
660 and location) and their temporal clustering score. Conceptually, a participant's memory
661 fingerprint describes their tendency to order, in their recall sequences (and presumably,
662 organize in memory), the studied words along each dimension. If these memory fin-
663 gerprints are truly unique to each participant, then we would expect that the estimated
664 fingerprints computed for a given participant, on different lists, should be more similar
665 than the estimated fingerprints computed for different participants. We reasoned that the
666 feature-rich condition would provide the best opportunity to test this assumption, since
667 the clustering scores would not be potentially confounded by order manipulations. To test

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

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

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

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

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

Beyond affecting participants’ memories for individual lists, we wondered how studying early lists (with versus without incidental visual features) might affect how participants remembered later lists (again, with versus without incidental visual features). In other words, we sought to test for potential effects of changing the “richness” of participants’

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

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

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

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

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 (Tab. 2) and reduced (Tab. 3) conditions. Participants in both conditions remembered more words on early versus late lists. Participants in the feature-rich (but not reduced) condition exhibited more temporal clustering on early versus late lists. And participants in both conditions tended to exhibit more semantic clustering on early versus late lists. Participants in the reduced (but not feature-rich) conditions tended to exhibit more lexicographic clustering on early versus late lists. Taken together, these comparisons suggest that even when the presence or absence of incidental visual features is stable across lists, participants still exhibit some differences in their performance and memory organization tendencies for early versus late lists.

With these differences in mind, we next compared participants' memories on early versus late lists for two additional experimental conditions (see *Random conditions*, Fig. S1).

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

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

696 In a *reduced (early)* condition, we held the visual features constant on early lists, but
697 allowed them to vary randomly on late lists. In a *reduced (late)* condition, we allowed
698 the visual features to vary randomly on early lists, but held them constant on late lists.
699 Given our above findings that (a) participants tended to exhibit stronger clustering effects
700 on feature-rich (versus reduced) lists, and (b) participants tended to remember more
701 words and exhibit stronger clustering effects on early (versus late) lists, we expected
702 these early versus late differences to be enhanced in the reduced (early) condition and
703 diminished in the reduced (late) condition. However, to our surprise, participants in *neither*
704 condition exhibited reliable early-versus-late differences in accuracy, temporal clustering,
705 nor feature-based clustering (Tabs. 4, 5). We hypothesized that adding or removing the
706 variability in the visual features was acting as a sort of “event boundary” between early
707 and late lists (e.g., Clewett et al., 2019; Radvansky and Copeland, 2006; Radvansky and
708 Zacks, 2017). In prior work, we (and others) have found that memories formed just
709 after event boundaries can be enhanced (e.g., due to less contextual interference between
710 pre- and post-boundary items; Flores et al., 2017; Gold et al., 2017; Manning et al., 2016;
711 Pettijohn et al., 2016).

712 We found that *adding* incidental visual features on later lists that had not been present
713 on early lists (as in the reduced (early) condition) served to enhance recall performance
714 relative to conditions where all lists had the same blends of features (Tabs. 6, 7; also

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

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

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

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

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

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

715 see Fig. S3A). However, *subtracting* irrelevant visual features on later lists that *had* been
716 present on early lists (as in the reduced (late) condition) did not appear to impact recall
717 performance (Tabs. 8, 9) These comparisons suggest that recall accuracy has a directional
718 component: accuracy is affected differently by removing features that had initially been
719 present versus adding features that had initially been absent. In contrast, we found
720 that participants exhibited more temporal and feature-based clustering when we added
721 incidental visual features to *any* lists (feature-rich versus reduced: Tab. 1; reduced versus
722 reduced (early): Tab. 7; reduced versus reduced (late): Tab. 9). Temporal and feature-based
723 clustering were not reliably different in the feature-rich versus reduced (early) or reduced
724 (late) conditions (Tabs. 6, 8).

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

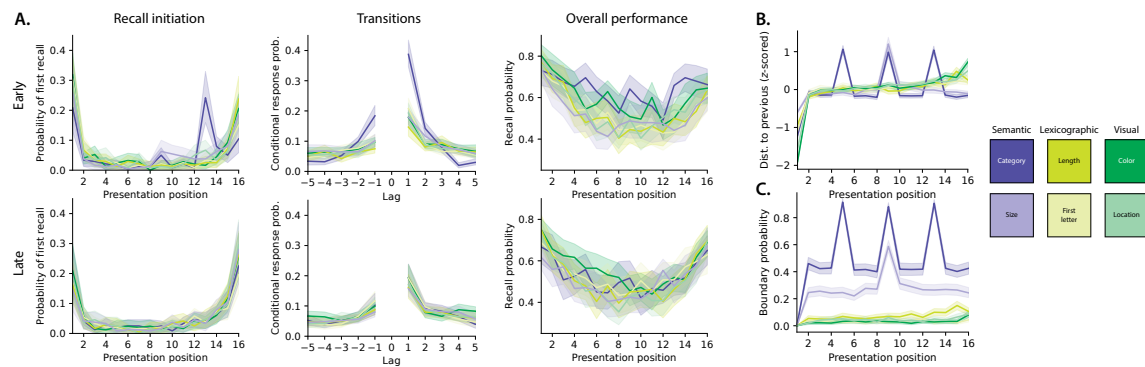


Figure 3: Recall dynamics in feature-rich free recall (order manipulation conditions). **A.** Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling each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

conditions, where variations in visual features were added or removed from a subset of the lists participants studied).

Across each of six order manipulation conditions, we sorted early lists by one feature dimension but randomly ordered the items on late lists (see *Order manipulation conditions*; features: category, size, length, first letter, color, and location). When we compared participants’ memories for early lists in each of these conditions to their memories for early lists in the feature-rich condition (Tab. 10), we found that participants in the category-ordered condition remembered more words than participants in the feature-rich condition. Participants’ performances on early lists in all of the other order manipulation conditions were indistinguishable from performance on the early feature-rich lists. We also compared

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

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

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

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

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

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

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

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

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

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

750 participants' temporal clustering on early lists in each of these conditions to their tem-
 751 poral clustering on early lists in the feature-rich condition (Tab. 11). Participants in both
 752 of the semantically ordered conditions and the first letter-ordered condition all exhibited
 753 stronger temporal clustering on early lists (vs. early feature-rich lists). Participants in the
 754 other order manipulation conditions all showed similar temporal clustering on early lists
 755 relative to early feature-rich lists. We also also compared feature-based clustering on early
 756 lists across the order manipulation and feature-rich conditions. Since these results were
 757 similar across both semantic conditions (category and size; Tab. 12), both lexicographic
 758 conditions (length and first letter; Tab. 13), and both visual conditions (color and loca-
 759 tion; Tab. 14), here we aggregate data from conditions that manipulated each of these
 760 three feature groupings in our comparisons, to simplify the presentation. On early lists,
 761 participants in the semantically ordered conditions exhibited stronger semantic clustering
 762 relative to participants in the feature-rich condition, but showed no reliable differences in
 763 lexicographic or visual clustering. Similarly, participants in the lexicographically ordered
 764 conditions exhibited stronger (relative to feature rich participants) lexicographic cluster-
 765 ing on early lists, but showed no reliable differences in semantic or visual clustering. And
 766 participants in the visually ordered conditions exhibited stronger visual clustering (again,
 767 relative to feature-rich participants, and on early lists), but showed no reliable differences
 768 in semantic or lexicographic clustering. Taken together, these order manipulation results

769 suggest several broad patterns (Figs. 3A, 4). First, most of the order manipulations we
770 carried out did *not* reliably affect overall recall performance. Second, most of the order
771 manipulations increased participants' tendencies to temporally cluster their recalls. Third,
772 all of the order manipulations enhanced participants' clustering of each condition's target
773 feature (i.e., semantic manipulations enhanced semantic clustering, lexicographic manip-
774 ulations enhanced lexicographic clustering, and visual manipulations enhanced visual
775 clustering; Fig. 5C) while leaving clustering along other feature dimensions roughly un-
776 changed (i.e., semantic manipulations did not affect lexicographic or visual clustering, and
777 so on). Although it is not possible to fully separate feature-based versus temporal clus-
778 tering when considering sorted lists, we used a permutation-based procedure to identify
779 the degree of feature clustering over and above what could be accounted for by temporal
780 clustering alone (see *Factoring out the effects of temporal clustering*). When we carried out
781 this analysis (Fig. 5D), we found that participants exhibited more semantic clustering on
782 semantically sorted lists than on randomly ordered lists, but the effects of the other order
783 manipulations could not reliably be separated from temporal clustering alone (reliable
784 comparisons are reported in the figure).

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

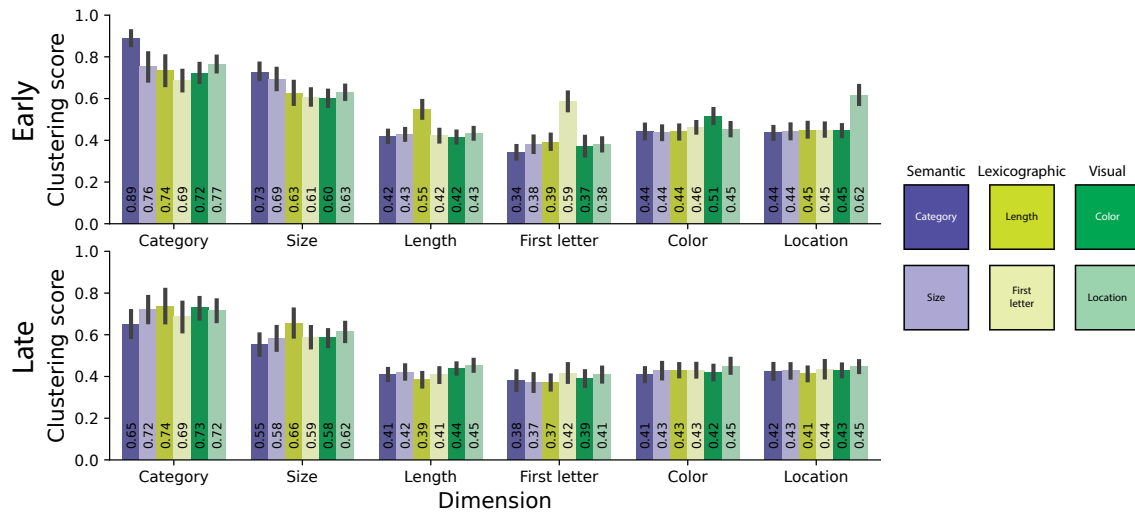


Figure 4: Memory “fingerprints” (order manipulation conditions). The across-participant average clustering scores for each feature type (*x*-axis) are displayed for each experimental condition (color), separately for order-manipulated (early, top) and randomly ordered (late, bottom) lists. Error bars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

794 *event boundaries*) with the positions of items participants recalled first, we noticed a strik-

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

796 the category labels changed every four items on average (dark purple peaks in Figs. 3B,

797 C), and participants also seemed to display an increased tendency (relative to other or-

798 der manipulation and random conditions) to initiate recall of category-ordered lists with

799 items whose study positions were integer multiples of four. Similarly, for size-ordered

800 lists, the size labels changed every eight items on average (light purple peaks in Figs. 3B,

801 C), and participants also seemed to display an increased tendency to initiate recall of

802 size-ordered lists with items whose study positions were integer multiples of eight. A

803 second striking difference is that participants in the category condition exhibited a much

804 steeper lag-CRP (Fig. 3A, top middle panel) than participants in other conditions. (This is

805 another expression of participants’ increased tendencies to temporally cluster their recalls

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

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

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

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

on category-ordered lists, as we reported above.) Taken together, these order-specific idiosyncrasies suggest a hierarchical set of influences on participants' memories. At longer timescales, "event boundaries" (to use the term loosely) can be induced across lists by adding or removing incidental visual features. At shorter timescales, "event boundaries" can be induced across items (within a single list) by adjusting how item features change throughout the list.

The above comparisons between memory performance on early lists in the order manipulation and feature-rich conditions highlight how sorted lists are remembered differently from random lists. We also wondered how sorting lists along each feature dimension influenced memory relative to sorting lists along the other feature dimensions (accuracy: Tab. 15; temporal clustering: Tab. 16; feature-based clustering: Tab 17). Participants trended towards remembering early lists that were sorted visually better than lexicographically sorted lists. However, participants showed no reliable differences in recall for semantically versus lexicographically or visually sorted lists. Participants temporally clustered semantically sorted lists more strongly than lists sorted either lexicographically or

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

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

visually, but did not show reliable differences in temporal clustering on lexicographically versus visually sorted lists. Participants also showed reliably more semantic clustering on semantically sorted lists than lexicographically or visually sorted lists; more lexicographic clustering on lexicographically sorted lists than semantically or visually sorted lists; and more visual clustering on visually sorted lists than semantically or lexicographically sorted lists. In summary, sorting lists by different features appeared to have slightly different effects on overall memory performance and temporal clustering. Participants also tended to cluster their recalls along a given feature dimension more when the studied lists were (versus were not) sorted along that dimension.

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

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

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

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

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

individuals).

At the group level, there seemed to be almost no lingering impact of sorting early lists on memory for later lists. To simplify the presentation, we report these null results in aggregate across the three feature groupings (accuracy: Tab. 18; temporal clustering: Tab. 19; feature-based clustering: Tabs. 20, 21, and 22). Relative to memory performance on late feature-rich lists, participants' memory performance in all six order manipulation conditions showed no reliable differences. Nor did we observe any reliable differences in temporal clustering on late lists (relative to late feature-rich lists). Aside from a slightly increased tendency for participants to cluster words by their length on late visual order manipulation lists (more than late feature-rich lists), we observed no reliable differences in any type of feature clustering on late order manipulation condition lists versus late feature-rich lists.

We also looked for more subtle group-level patterns. For example, perhaps sorting early lists by one feature dimension could affect how participants cluster *other* features (on early and/or late lists) as well. As described above, a participant's memory finger-

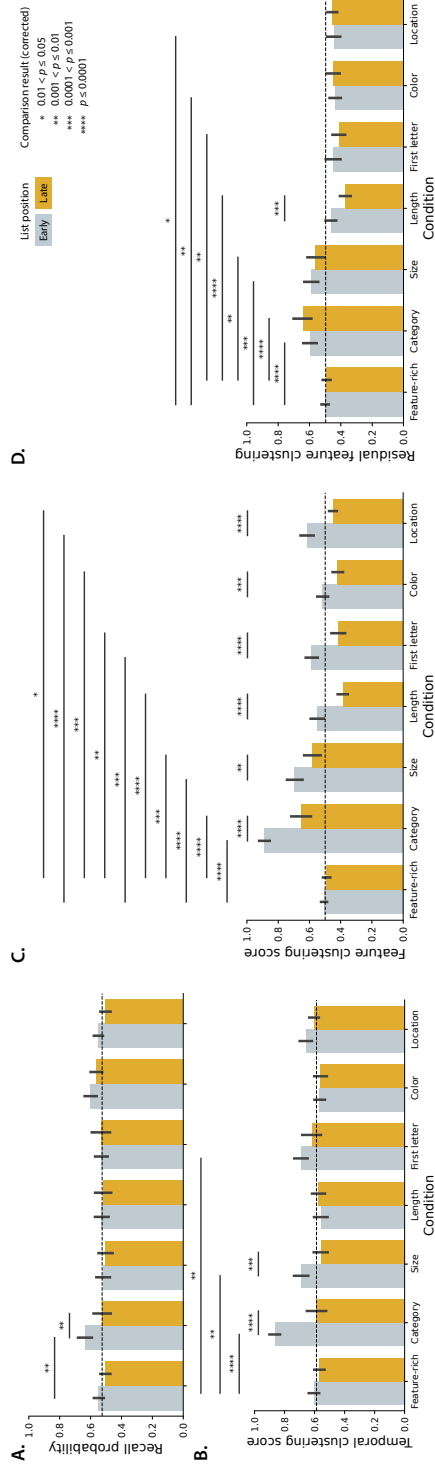


Figure 5: Recall probability and clustering scores on early and late lists. The bar heights display the average (across participants) recall probabilities (A.), temporal clustering scores (B.), feature clustering scores (C.), and residual feature clustering scores (after factoring out temporal clustering effects; D.) for early (gray) and late (gold) lists. For the feature-rich bars (left), the feature clustering scores are averaged across all feature dimensions. For the order manipulation conditions, feature clustering scores are displayed for the focused-on feature for each condition (e.g., category clustering scores are displayed for the category condition, and so on). All panels: error bars denote bootstrap-estimated 95% confidence intervals. The horizontal dotted lines denote the average values (across all lists and participants) for the feature rich condition. The bars denote t -tests between the corresponding bars, and the asterisks denote the Benjamini-Hochberg-corrected p -values. Comparisons for which corrected $p \geq 0.05$ are not shown.

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

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

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

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

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

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

print characterizes how they tend to retrieve memories of the studied items, perhaps searching in parallel through several feature spaces (or along several representational dimensions). To gain insights into the dynamics of how participants' clustering scores tended to change over time, we computed the average (across participants) fingerprint from each list, from each order manipulation condition (Fig. 6). We projected these fingerprints into a two-dimensional space to help visualize the dynamics (top panels; see *Computing low-dimensional embeddings of memory fingerprints*). We found that participants' average fingerprints tended to remain relatively stable on early lists, and exhibited a "jump" to another stable state on later lists. The sizes of these jumps varied somewhat across conditions (the Euclidean distances between fingerprints in their original high dimensional spaces are displayed in the bottom panels). We also averaged the fingerprints across early and late lists, respectively, for each condition (Fig. 6B). We found that participants' fingerprints on early lists seem to be influenced by the order manipulations for those lists (see the locations of the circles in Fig. 6B). There also seemed to be some consistency across different features within a broader type. For example, both semantic feature conditions (category and size; purple markers) diverge in a similar direction from the group; both lexicographic feature conditions (length and first letter; yellow markers) diverge in a similar direction; and both visual conditions (color and location; green) also diverge in a similar direction. But on late lists, participants' fingerprints seem to return to a common state that is roughly shared across conditions (i.e., the stars in that panel are clumped together).

When we examined the data at the level of individual participants (Figs. 7 and 8), a clearer story emerged. Within each order manipulation condition, participants exhibited a range of feature clustering scores on both early and late lists (Fig. 7A, B). Across every order manipulation condition, participants who exhibited stronger feature clustering

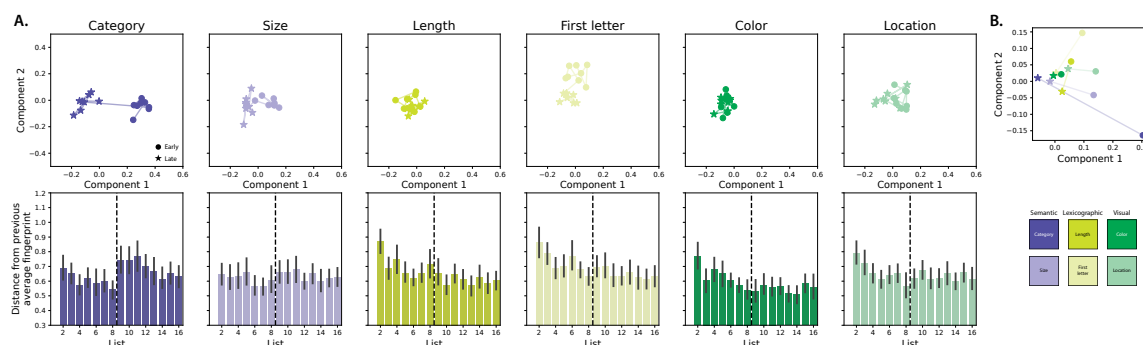


Figure 6: Memory fingerprint dynamics (order manipulation conditions). **A.** Each column (and color) reflects an experimental condition. In the top panels, each marker displays a 2D projection of the (across-participant) average memory fingerprint for a single list. Lines connect successive lists. Order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

876 (for their condition's manipulated feature) recalled more words. This trend held overall
877 across conditions and participants (early: $r(179) = 0.492$, $p < 0.001$, $CI = [0.352, 0.606]$;
878 late: $r(179) = 0.403$, $p < 0.001$, $CI = [0.271, 0.517]$) as well as for each condition indi-
879 vidually for early ($r_s \geq 0.331$, all $p_s \leq 0.069$) and late ($r_s \geq 0.404$, all $p_s \leq 0.027$) lists.
880 We found no evidence of a condition-level trend; for example, the conditions where
881 participants tended to show stronger clustering scores were not correlated with the con-
882 ditions where participants remembered more words (early: $r(4) = 0.511$, $p = 0.300$, $CI =$
883 $[-0.999, 0.996]$; late: $r(4) = -0.304$, $p = 0.559$, $CI = [-0.833, 0.748]$; see insets of Fig. 7A
884 and B). We observed carryover associations between feature clustering and recall perfor-
885 mance (Fig. 7C, D). Participants who showed stronger feature clustering on early lists
886 in the non-visual conditions tended to recall more items on late lists (across conditions:
887 $r(179) = 0.230$, $p = 0.002$, $CI = [0.072, 0.372]$; all non-visual conditions individually: r_s

888 ≥ 0.405 , all $ps \leq 0.027$; color: $r(29) = 0.212$, $p = 0.251$, $CI = [-0.164, 0.532]$; location:
 889 $r(28) = 0.320$, $p = 0.085$, $CI = [0.011, 0.584]$). Participants who recalled more items on
 890 early lists also tended to show stronger feature clustering on late lists (across conditions:
 891 $r(179) = 0.464$, $p < 0.001$, $CI = [0.321, 0.582]$; individual conditions: all $rs \geq 0.377$, all ps
 892 ≤ 0.040). Neither of these effects showed condition-level trends (early feature clustering
 893 versus late recall probability: $r(4) = -0.338$, $p = 0.512$, $CI = [-0.971, 0.634]$; early recall
 894 probability versus late feature clustering: $r(4) = 0.451$, $p = 0.369$, $CI = [-0.986, 0.998]$). We
 895 also looked for associations between feature clustering and temporal clustering. Across
 896 every order manipulation condition, participants who exhibited stronger feature cluster-
 897 ing also exhibited stronger temporal clustering. For early lists (Fig. 7E), this trend held
 898 overall ($r(179) = 0.916$, $p < 0.001$, $CI = [0.893, 0.936]$), for each condition individually
 899 (all $rs \geq 0.822$, all $ps < 0.001$), and across conditions ($r(4) = 0.964$, $p = 0.002$). For late
 900 lists (Fig. 7F), the results were more variable (overall: $r(179) = 0.348$, $p < 0.001$; all non-
 901 visual conditions: $rs \geq 0.382$, all $ps \leq 0.037$; color: $r(29) = 0.453$, $p = 0.011$; location:
 902 $r(28) = 0.190$, $p = 0.314$; across-conditions: $r(4) = -0.036$, $p = 0.945$). While less ro-
 903 bust than the carryover associations between feature clustering and recall performance,
 904 we also observed some carryover associations between feature clustering and temporal
 905 clustering (Fig. 7G, H). Participants who showed stronger feature clustering on early lists
 906 showed stronger temporal clustering on later lists (overall: $r(179) = 0.464$, $p < 0.001$, $CI =$
 907 $[0.321, 0.582]$; for individual conditions: all $rs \geq 0.377$, all $ps \leq 0.040$; across conditions:
 908 $r(4) = 0.451$, $p = 0.369$, $CI = [-0.986, 0.998]$). And participants who showed stronger
 909 temporal clustering on early lists trended towards showing stronger feature clustering on
 910 later lists (overall: $r(179) = 0.266$, $p < 0.001$, $CI = [0.129, 0.396]$; for individual conditions:
 911 all $rs \geq 0.298$, all $ps \leq 0.110$; across conditions: $r(4) = 0.064$, $p = 0.903$, $CI = [-0.972, 0.965]$.
 912 Taken together, the results displayed in Figure 7 show that participants who were more

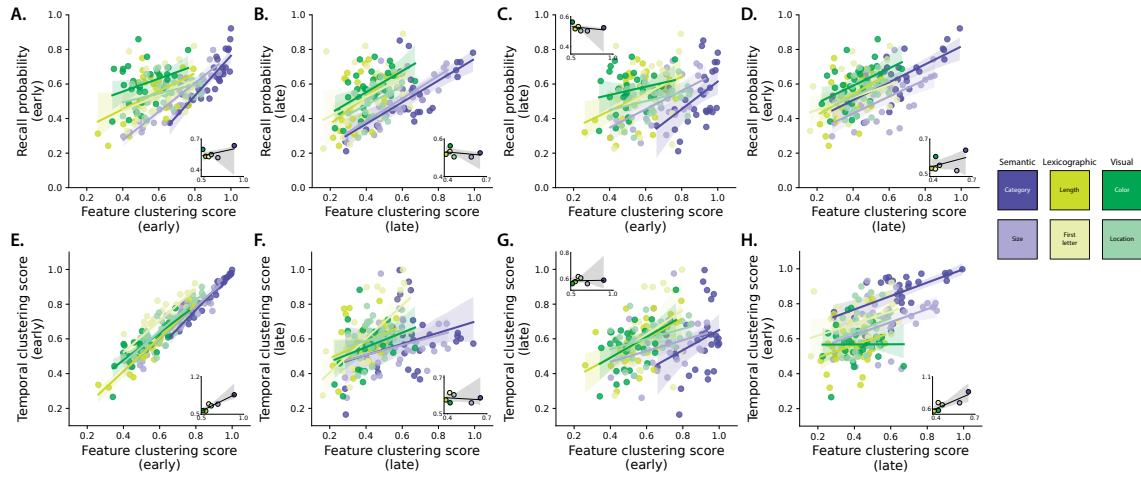


Figure 7: Interactions between feature clustering, recall probability, and contiguity. **A.** Recall probability versus feature clustering scores for order manipulation (early) lists. **B.** Recall probability versus feature clustering for randomly ordered (late) lists. **C.** Recall probability on late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

913 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
 914 stronger temporal clustering. These associations also appeared to carry over to later lists,
 915 even when the items on those later lists were presented in a random order.

917 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
 918 clustering on early lists often tended to show strong feature clustering on late lists (Fig. 8A;
 919 overall across participants and conditions: $r(179) = 0.591$, $p < 0.001$, $CI = [0.472, 0.682]$;
 920 category: $r(28) = 0.590$, $p < 0.001$, $CI = [0.354, 0.756]$; size: $r(28) = 0.488$, $p =$
 921

922 0.006, CI = [0.134, 0.732]; length: $r(28) = 0.384$, $p = 0.036$, CI = [0.040, 0.681]; first
 923 letter: $r(28) = 0.202$, $p = 0.284$, CI = [-0.273, 0.620]; color: $r(29) = -0.183$, $p = 0.325$, CI =
 924 [-0.562, 0.258]; location: $r(28) = 0.031$, $p = 0.870$, CI = [-0.240, 0.296]; across conditions:
 925 $r(4) = 0.942$, $p = 0.005$, CI = [0.442, 1.000]). Although participants tended to show weaker
 926 feature clustering on late lists (Figs. 4, 5) on *average*, the associations between early and late
 927 lists for individual participants suggests that some influence of early order manipulations
 928 may linger on late lists. We found that participants who exhibited larger carryover in
 929 feature clustering (i.e., continued to show strong feature clustering on late lists) for the
 930 semantic order manipulations (but not other manipulations) also tended to show a smaller
 931 decrease in recall on early versus late lists (Fig. 8B; overall: $r(179) = 0.307$, $p < 0.001$, CI =
 932 [0.148, 0.469]; category: $r(28) = 0.350$, $p = 0.058$, CI = [0.050, 0.642]; size: $r(28) =$
 933 0.708 , $p < 0.001$, CI = [0.472, 0.862]; length: $r(28) = 0.205$, $p = 0.276$, CI = [-0.109, 0.492];
 934 first letter: $r(28) = 0.081$, $p = 0.672$, CI = [-0.433, 0.597]; color: $r(29) = 0.155$, $p =$
 935 0.406 , CI = [-0.174, 0.541]; location: $r(28) = 0.052$, $p = 0.787$, CI = [-0.307, 0.360]; across
 936 conditions: $r(4) = 0.635$, $p = 0.176$, CI = [-0.924, 0.981]). Participants who exhibited
 937 larger carryover in feature clustering also tended to show stronger temporal clustering
 938 on late lists (relative to early lists) for all but the category condition (Fig. 8C; overall:
 939 $r(179) = 0.426$, $p < 0.001$, CI = [0.285, 0.544]; category: $r(28) = 0.110$, $p = 0.564$, CI =
 940 [-0.284, 0.442]; all non-category conditions: all r s ≥ 0.406 , all p s ≤ 0.023 ; across conditions:
 941 $r(4) = 0.649$, $p = 0.163$, CI = [-0.856, 0.988]).

942 We suggest two potential interpretations of these findings. First, it is possible that
 943 some participants are more “malleable” or “adaptable” with respect to how they organize
 944 incoming information. When presented with list of items sorted along *any* feature dimen-
 945 sion, they will simply adopt that feature as a dominant dimension for organizing those
 946 items and subsequent (randomly ordered) items. This flexibility in memory organization

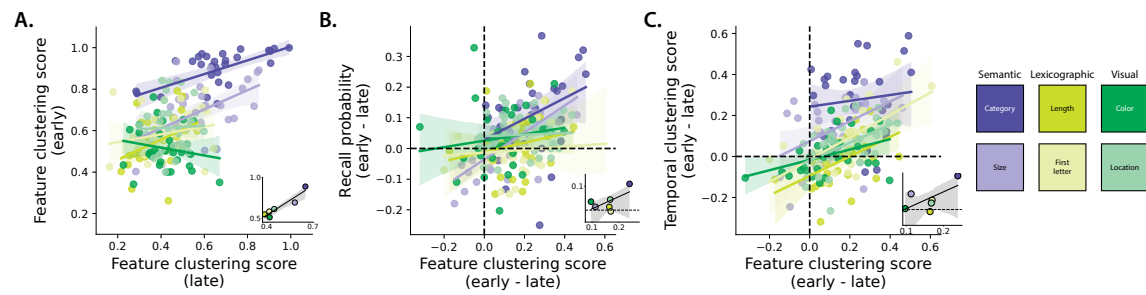


Figure 8: Feature clustering carryover effects. **A.** Feature clustering scores for order manipulation (early) versus randomly ordered (late) lists. **B.** Accuracy differences (on early versus late lists) versus feature clustering “carryover” (defined as the differences between the average clustering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

might afford such participants a memory advantage, explaining their strong recall performance. An alternative interpretation is that each participant comes into our study with a “preferred” way of organizing incoming information. If they happen to be assigned to an order manipulation condition that matches their preferences, then they will appear to be “sensitive” to the order manipulation and also exhibit a high degree of carryover in feature clustering from early to late lists. These 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” anal-*

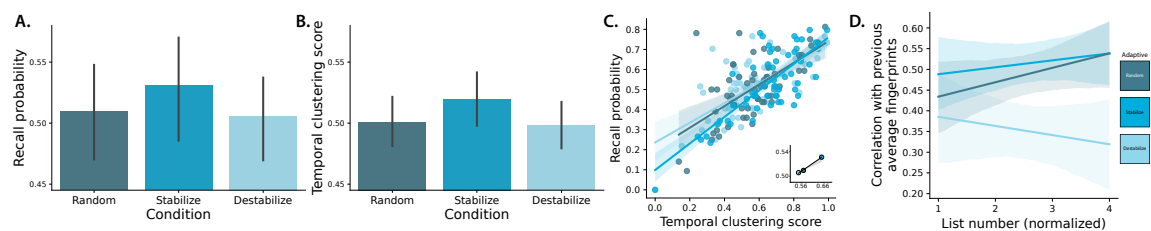


Figure 9: Adaptive free recall. **A.** Average probability of recall (taken across words, lists, and participants) for each batch of four lists in the adaptive condition. **B.** Average temporal clustering scores for each batch of lists. **C.** Recall probability versus temporal clustering scores by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

ysis). Third, on *destabilize* lists, the presentation order is adjusted to be *minimally* similar to the current estimate of the participant’s memory fingerprint (see *Ordering “stabilize” and “destabilize” lists by an estimated fingerprint*). The orders in which participants experienced each type of list were counterbalanced across participants to help reduce the influence of potential list-order effects. Because the presentation orders on stabilize and destabilize lists are adjusted to best match each participant’s personal memory fingerprint, the adaptive condition removes uncertainty about whether participants’ assigned conditions might just “happen” to match their preferred ways of organizing their memories.

Participants’ fingerprints on stabilize and random lists tended to become (numerically) slightly more similar to their average fingerprints computed from the previous lists they had experienced, and their fingerprints on destabilize lists tended to become numerically less similar (Fig. 9D). Overall, we found that participants tended to be better at remembering words on stabilize lists relative to words on both random ($t(59) = 1.740$, $p = 0.087$, $d = 0.095$, $CI = [-0.187, 3.761]$) and destabilize ($t(59) = 1.714$, $p = 0.092$, $d = 0.114$, $CI = [-0.351, 4.108]$) lists (Fig. 9A). Participants showed no reliable differences in their memory

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

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

performance on destabilize versus random lists ($t(59) = -0.249, p = 0.804, d = -0.017, CI = [-2.327, 1.578]$). Participants also exhibited stronger temporal clustering on stabilize lists, relative to random ($t(59) = 3.428, p = 0.001, d = 0.306, CI = [1.635, 5.460]$) and destabilize ($t(59) = 4.174, p < 0.001, d = 0.374, CI = [1.964, 6.968]$) lists (Fig. 9B). We found no reliable differences in temporal clustering for items on random versus destabilize lists ($t(59) = -0.880, p = 0.382, d = -0.081, CI = [-3.165, 1.127]$).

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

We note that the multivariate aspect of the adaptive condition (i.e., sorting lists simultaneously along multiple feature dimensions) provides an important contrast with

the other order manipulation conditions, where we sort lists along only a single feature dimension in each condition. We found that participants “naturally” clustered their recalls along multiple feature dimensions, even when the lists they studied were not sorted along those dimensions (as in the feature-rich condition). A caveat is that the *specific* feature dimensions participants tended to cluster along varied across participants. One way to quantify the multidimensional nature of participants’ clustering tendencies is to sort each participant’s clustering scores (for each of the six feature dimensions, along with a seventh dimension to capture temporal clustering). We can then ask whether the distribution of clustering scores at each “rank” within the sorted set of scores for each participant has a mean that is reliably different from a chance value of 0.5. We carried out these tests for each set of ranked scores, and found that participants in the feature-rich condition reliably cluster their recalls along at least three dimensions, including temporal clustering (which was often ranked highest; Tab. 23). That the clustering scores ranked in the top three dimensions were reliably greater than chance suggests that participants organize their memories along at least three feature dimensions, even when the words are studied in a random order.

Discussion

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

Recap: visual feature manipulations

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

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

1042 more tightly associated with each other.

1043 When we held the visual features constant during early lists, but then varied words'
1044 appearances in later lists (i.e., the reduced (early) condition), participants' overall memory
1045 performance improved. However, this impact was directional: when we *removed* visual
1046 features from words in late lists that had been present in early lists (i.e., the reduced (late)
1047 condition), we saw no memory improvement.

1048 **Recap: order manipulations**

1049 When we (stochastically) sorted early lists along different feature dimensions, we found
1050 several impacts on participants' memories. Sorting early lists semantically (by word cat-
1051 egory) enhanced participants' memories for those lists, but the effects on performance of
1052 sorting along other feature dimensions were inconclusive. However, each order manipu-
1053 lation substantially affected how participants *organized* their memories of words from the
1054 ordered lists. When we sorted lists semantically, participants displayed stronger semantic
1055 clustering; when we sorted lists lexicographically, they displayed stronger lexicographic
1056 clustering; and when we sorted lists visually, they displayed stronger visual clustering.
1057 Clustering along the unmanipulated feature dimensions in each of these cases was un-
1058 changed.

1059 The order manipulations we examined also appeared to induce, in some cases, a
1060 tendency to “clump” similar words within a list. This was most apparent on semantically
1061 ordered lists, where the probability of initiating recall with a given word seemed to follow
1062 groupings defined by feature change points.

1063 We also examined the impact of early list order manipulations on memory for late
1064 lists. At the group level, we found little evidence for lingering “carryover” effects of
1065 these manipulations: participants in the order manipulation conditions showed similar

memory performance and clustering on late lists to participants in the corresponding control (feature-rich) condition. At the level of individual participants, however, we found several meaningful patterns.

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

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

gerprints, they remembered more words overall and showed stronger temporal clustering. This comports well with the second possibility described above. Specifically, each participant seems to bring into the experiment their own idiosyncratic preferences and strategies for organizing the words in their memory. When we presented the words in an order consistent with each participant's idiosyncratic fingerprint, their memory performance improved. This might indicate that the participants were spending less cognitive effort "reorganizing" the incoming words on those lists, which freed up resources to devote to encoding processes instead.

Memory consequences of feature variability

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

Some prior studies have suggested that people can "cue" their memories using different "strategies" or "pathways" for searching for the target information. For example, modern accounts of free recall often posit that memory search typically begins by matching

1115 the current state of mental context with the contexts associated with other items in mem-
1116 ory (Kahana, 2020). Since context is the defining hallmark of episodic memory (Tulving,
1117 1983), context-based search can be described as an “episodic” pathway to recall. When
1118 episodic cueing fails to elicit a match, participants may then search for items that are simi-
1119 lar to the current mental context or mental state along other dimensions, such as semantic
1120 similarity (Davachi et al., 2003; Socher et al., 2009). These multiple pathways accounts of
1121 memory search also provide a potential explanation of why participants might have an
1122 easier time remembering richer stimuli (or experiences): richer stimuli and experiences
1123 might have more features that could be used to cue memory search. Our work suggests
1124 that there may be some additional factors at play with respect to the *dynamics* of these pro-
1125 cesses. In particular, we only observed memory benefits for “richer” stimuli when they
1126 were encountered after more “impoverished” stimuli (in the reduced (early) condition).
1127 This suggests that the pathways available to recall a given item may also depend on recent
1128 prior experiences.

1129 We did *not* find any evidence that changing words’ appearances *harmed* memory per-
1130 formance, e.g., by distracting them with irrelevant information (Lange, 2005; Marsh et al.,
1131 2012, 2015; Reinitz et al., 1992). Nor did we find any evidence that *changes* in the presence
1132 of potentially “distracting” features adversely affected memory. For example, when we
1133 increased or decreased the variability in words’ appearances on late versus early lists (as in
1134 the reduced (early) and reduced (late) conditions), we found no evidence that this harmed
1135 participants’ memories. One potential interpretation under the “multiple pathways to
1136 recall” framework is that the availability of multiple pathways to recall do not appear to
1137 specifically interfere with each other.

1138 **Context effects on memory performance and organization**

1139 In real-world experiences, each moment's unique blend of contextual features (where we
1140 are, who we are with, what else we are thinking of at the time, what else we experience
1141 nearby in time, etc.) plays an important role in how we interpret, experience, and re-
1142 member that moment, and how we relate it to our other experiences (e.g., for review see
1143 Manning, 2020). What are the analogues of real-world contexts in laboratory tasks like
1144 the free recall paradigm employed in our study? In general, modern formal accounts of
1145 free recall (Kahana, 2020) describe context as comprising a mix of (a) features pertaining
1146 to or associated with each item and (b) other items and thoughts experienced nearby in
1147 time, e.g., that might still be "lingering" in the participant's thoughts at the time they
1148 study the item. Item features can include semantic properties (i.e., features related to the
1149 item's meaning), lexicographic properties (i.e., features related to the item's letters), sen-
1150 sory properties (i.e., feature related to the item's appearance, sound, smell, etc.), emotional
1151 properties (i.e., features related to how meaningful the item is, whether the item evokes
1152 positive or negative feelings, etc.), utility-related properties (e.g., features that describe
1153 how an item might be used or incorporated into a particular task or situation), and more.
1154 Essentially any aspect of the participant's experience that can be characterized, measured,
1155 or otherwise described can be considered to influence the participant's mental context at
1156 the moment they experience that item. Temporally proximal features include aspects of
1157 the participant's internal or external experience that are *not* specifically occurring at the
1158 moment they encounter an item, but that nonetheless influence how they process the item.
1159 Thoughts related to percepts, goals, expectations, other experiences, and so on that might
1160 have been cued (directly or indirectly) by the participant's recent experiences prior to the
1161 current moment all fall into this category. Internally driven mental states, such as thinking
1162 about an experience unrelated to the experiment, also fall into this category.

1163 Contextual features need not be intentionally or consciously perceived by the partic-
1164 ipant to affect memory, nor do they need to be relevant to the task instructions or the
1165 participant’s goals. Incidental factors such as font color (Jones and Pyc, 2014), back-
1166 ground color (Isarida and Isarida, 2007), inter-stimulus images (Chiu et al., 2021; Ger-
1167 shman et al., 2013; Manning et al., 2016), background sounds (Beaman and Jones, 1998;
1168 Sahakyan and Smith, 2014), secondary tasks (Masicampo and Sahakyan, 2014; Oberauer
1169 and Lewandowsky, 2008; Polyn et al., 2009), and more can all impact how participants
1170 remember, and organize in memory, lists of studied items.

1171 Consistent with this prior work, we found that participants were sensitive to task-
1172 irrelevant visual features. We also found that changing the dynamics of those task-
1173 irrelevant visual features (in the reduced (early) and reduced (late) conditions) *also* affected
1174 participants’ memories. This suggests that it is not only the contextual features themselves
1175 that affect memory, but also the *dynamics* of context—i.e., how the contextual features
1176 associated with each item change over time.

1177 **Priming effects on memory performance and organization**

1178 When our ongoing experiences are ambiguous, we can draw on our past experiences,
1179 expectations, and other real, perceived, or inferred cues to help resolve these ambiguities.
1180 We may also be overtly or covertly “primed” to influence how we are likely to resolve
1181 ambiguities. For example, before listening to a story with several equally plausible inter-
1182 pretations, providing participants with “background” information beforehand can lead
1183 them towards one interpretation versus another (Yeshurun et al., 2017). More broadly, our
1184 conscious and unconscious biases and preferences can influence not only how we interpret
1185 high-level ambiguities, but even how we process low-level sensory information (Katabi
1186 et al., 2023).

1187 In more simplified scenarios, like list-learning paradigms, the stimuli and tasks partic-
1188 ipants encounter before studying a given list can influence what and how they remember.
1189 For example, when participants are directed to suppress, disregard, or ignore “distracting”
1190 stimuli early on in an experiment, participants often tend to remember those stimuli less
1191 well when they are re-used as to-be-remembered targets later on in the experiment (Tip-
1192 per, 1985). In general, participants’ memories can be influenced by exposing them to
1193 a wide range of positive and negative priming factors before they encounter the to-be-
1194 remembered information (Balota et al., 1992; Clayton and Chattin, 1989; Donnelly, 1988;
1195 Flexser and Tulving, 1982; Gotts et al., 2012; Huang et al., 2004; Huber, 2008; Huber et al.,
1196 2001; McNamara, 1994; Neely, 1977; Rabinowitz, 1986; Tulving and Schacter, 1991; Watkins
1197 et al., 1992; Wiggs and Martin, 1998).

1198 The order manipulation conditions in our experiment show that participants can also be
1199 primed to pick up on more subtle statistical structure in their experiences, like the dynamics
1200 of how the presentation orders of stimuli vary along particular feature dimensions. These
1201 order manipulations affected not only how participants remembered the manipulated
1202 lists, but also how they remembered *future* lists with different (randomized) temporal
1203 properties.

1204 **Free recall of blocked versus random categorized word lists**

1205 A large number of prior studies have compared participants’ memories for categorized
1206 word lists that are presented in blocked versus random orders. In “blocked” lists, all
1207 of the words from a given semantic category (e.g., animals) are presented consecutively,
1208 whereas in “random” lists, the words from different categories are intermixed. Most of
1209 these studies report that participants tend to better remember blocked (versus random)
1210 lists (Bower et al., 1969; Cofer et al., 1966; D’Agostino, 1969; Dallett, 1964; Kintsch, 1970;

1211 Luek et al., 1971; Puff, 1974; Shapiro, 1970; Shapiro and Palermo, 1970). Other studies
1212 suggest that these order effects may also be modulated by factors like list length and the
1213 numbers of exemplars in each category (e.g., Borges and Mangler, 1972).

1214 Although we did not directly manipulate “blocking” in our order manipulation condi-
1215 tions, our sorting procedures in those conditions (see *Constructing feature-sorted lists*) have
1216 *indirect* effects on the lists’ blockiness. For example, lists that are stochastically sorted by
1217 semantic category will tend to contain runs of several same-category words in succession.
1218 Consistent with the above work on blocked versus random categorized lists, we found
1219 that participants tended to better remember lists that were sorted semantically (Fig. 5B).
1220 However, this memory improvement did not appear to extend to the other order ma-
1221 nipulation conditions we considered (e.g., to lexicographically or visually sorted lists).
1222 One possibility is that the memory benefits of blocked versus random lists are specific to
1223 semantic categories, and do not generalize to other feature dimensions. Another possi-
1224 bility is that the memory benefits are due to the presence of infrequent “jumps” between
1225 successive items (e.g., from different categories). Because the features we manipulated in
1226 the lexicographic and visual conditions were less categorical than the semantic features,
1227 feature values across words in those conditions tended to vary more gradually. Relatively
1228 stable features that are punctuated by infrequent large changes (e.g., as words transition
1229 from a same-category sequence to a new category) may also relate to perceived “event
1230 boundaries,” which can have important consequences for memory (DuBrow and Davachi,
1231 2013, 2016; DuBrow et al., 2017; Radvansky and Zacks, 2017).

1232 **Expectation, event boundaries, and situation models**

1233 Our findings that participants’ current and future memory behaviors are sensitive to
1234 manipulations in which features change over time, and how features change across items

1235 and lists, suggest parallels with studies on how we form expectations and predictions,
1236 segment our continuous experiences into discrete events, and make sense of different
1237 scenarios and situations. Each of these real-world cognitive phenomena entail identifying
1238 statistical regularities in our experiences, and exploiting those regularities to gain insight,
1239 form inferences, organize or interpret memories, and so on. Our past experiences enable
1240 us to predict what is likely to happen in the future, given what happened “next” in our
1241 previous experiences that were similar to now (Barron et al., 2020; Brigard, 2012; Chow
1242 et al., 2016; Eichenbaum and Fortin, 2009; Gluck et al., 2002; Goldstein et al., 2021; Griffiths
1243 and Steyvers, 2003; Jones and Pashler, 2007; Kim et al., 2014; Manning, 2020; Tamir and
1244 Thornton, 2018; Xu et al., 2023).

1245 When our expectations are violated, such as when our observations disagree with our
1246 predictions, we may perceive the “rules” or “situation” to have changed. *Event boundaries*
1247 denote abrupt changes in the state of our experience, for example, when we transition
1248 from one situation to another (Radvansky and Zacks, 2017; Zwaan and Radvansky, 1998).
1249 Crossing an event boundary can impair our memory for pre-boundary information and en-
1250 hance our memory for post-boundary information (DuBrow and Davachi, 2013; Manning
1251 et al., 2016; Radvansky and Copeland, 2006; Sahakyan and Kelley, 2002). Event bound-
1252 aries are also tightly associated with the notion of *situation models* and *schemas*—mental
1253 frameworks for organizing our understanding about the rules of how we and others are
1254 likely to behave, how events are likely to unfold over time, how different elements are
1255 likely to interact, and so on. For example, a situation model pertaining to a particular
1256 restaurant might set our expectations about what we are likely to experience when we
1257 visit that restaurant (e.g., what the building will look like, how it will smell when we enter,
1258 how crowded the restaurant is likely to be, the sounds we are likely to hear, etc.). Similarly,
1259 as mentioned in the *Introduction*, we might learn a schema describing how events are likely

1260 to unfold *across* any sit-down restaurant—e.g., open the door, wait to be seated, receive a
1261 menu, decide what to order, place the order, and so on. Situation models and schemas can
1262 help us to generalize across our experiences, and to generate expectations about how new
1263 experiences are likely to unfold. When those expectations are violated, we can perceive
1264 ourselves to have crossed into a new situation.

1265 In our study, we found that abruptly changing the “rules” about how the visual
1266 appearances of words are determined, or about the orders in which words are presented,
1267 can lead participants to behave similarly to what one might expect upon crossing an event
1268 boundary. Adding variability in font color and presentation location for words on late
1269 lists, after those visual features had been held constant on early lists, led participants to
1270 remember more words on those later lists. One potential explanation is that participants
1271 perceive an “event boundary” to have occurred when they encounter the first “late” list.
1272 According to contextual change accounts of memory across event boundaries (e.g., Flores
1273 et al., 2017; Gold et al., 2017; Pettijohn et al., 2016; Sahakyan and Kelley, 2002), this could
1274 help to explain why participants in the reduced (early) condition exhibited better overall
1275 memory performance. Specifically, their memory for late list items could benefit from less
1276 interference from early list items, and the contextual features associated with late list items
1277 (after the “event boundary”) might serve as more specific recall cues for those late items
1278 (relative to if the boundary had not occurred).

1279 **How do different types of clustering relate to each other, and to memory perfor-** 1280 **mance?**

1281 When the words on a studied list are presented in a random order, different types of
1282 clustering in participants’ recalls often tend to be negatively correlated. For example,
1283 words that occur nearby on the list will not (on average) tend to be semantically related, and

1284 vice versa. Therefore a participant who shows a strong tendency to temporally cluster their
1285 recalls will tend to show weaker semantic clustering, and so on (Healey and Uitvlugt, 2019;
1286 Howard and Kahana, 2002b; Sederberg et al., 2010). Further, there is some evidence that
1287 temporal clustering is positively correlated with memory performance, whereas semantic
1288 clustering is negatively correlated with memory performance (Sederberg et al., 2010).

1289 The notion of “multiple pathways to recall” discussed above (see *Memory consequences*
1290 *of feature variability*) suggests one potential explanation for these patterns. For exam-
1291 ple, temporal clustering has been proposed to reflect reliance on contextual cues in an
1292 “episodic” pathway to search memory, whereas semantic clustering reflects a relies on
1293 specific item features. These two pathways may “compete” with each other during re-
1294 call (Socher et al., 2009). Meanwhile, extra-list intrusion errors (i.e., false “recalls” of items
1295 that were never encountered on the list) often tend to share semantic features with recently
1296 recalled items (Zaromb et al., 2006) and also often lead the participant to stop recalling
1297 additional items (Miller et al., 2012). Speculatively, over-reliance on semantic cues may
1298 lead to more intrusion errors, which in turn may lead to fewer recalls overall.

1299 Our findings extend these prior results to consider lists that are *not* ordered randomly.
1300 Because ordering the words on a list along a particular feature dimension removes the
1301 “conflict” between temporal and feature clustering, the order manipulation conditions in
1302 our study represent an “edge case” whereby different pathways to recall are not neces-
1303 sarily in conflict with each other. For example, the same participants who exhibit strong
1304 feature clustering *also* show strong temporal clustering on ordered lists (Fig. 7E). This
1305 is presumably at least partly due to an inability to separate temporal and feature clus-
1306 tering on ordered lists (also see *Factoring out the effects of temporal clustering*). However,
1307 features that change gradually with time (i.e., presentation position) could also serve to
1308 strengthen the episodic (contextual) cues associated with each item. In other words, par-

1309 ticipants might essentially combine multiple noisy measures of change to form a more
1310 stable internal representation of temporal context.

1311 **Theoretical implications**

1312 Although most modern formal theories of episodic memory have been developed and
1313 tested to explain memory for list-learning tasks (Kahana, 2020), a number of recent studies
1314 suggest some substantial differences between memory for lists versus naturalistic stim-
1315 uli (e.g., real-world experiences, narratives, films, etc.; Heusser et al., 2021; Lee et al., 2020;
1316 Manning, 2021; Nastase et al., 2020). One reason is that naturalistic stimuli are often much
1317 more engaging than the highly simplified list-learning tasks typically employed in the
1318 psychological laboratory, perhaps leading participants to pay more attention, exert more
1319 effort, and stay more consistently motivated to perform well (Nastase et al., 2020). Another
1320 reason is that the temporal unfoldings of events and occurrences in naturalistic stimuli
1321 tend to be much more meaningful than the temporal unfoldings of items on typical lists
1322 used in laboratory memory tasks. Real-world events exhibit important associations at a
1323 broad range of timescales. For example, an early detail in a detective story may prove to
1324 be a clue to solving the mystery later on. Further, what happens in one moment typically
1325 carries some predictive information about what came before or after (Xu et al., 2023). In
1326 contrast, the lists used in laboratory memory tasks are most often ordered randomly, by
1327 design, to *remove* meaningful temporal structure in the stimulus (Kahana, 2012).

1328 On one hand, naturalistic stimuli provide a potential means of understanding how our
1329 memory systems function in the circumstances we most often encounter in our everyday
1330 lives. This implies that, to understand how memory works in the “real world,” we should
1331 study memory for stimuli that reflect the relevant statistical structure of real-world expe-
1332 riences. On the other hand, naturalistic stimuli can be difficult to precisely characterize or

1333 model, making it difficult to distinguish whether specific behavioral trends follow from
1334 fundamental workings of our memory systems, from some aspect of the stimulus, or from
1335 idiosyncratic interactions or interference between participants' memory systems and the
1336 stimulus. This challenge implies that, to understand the fundamental nature of memory
1337 in its "pure" form, we should study memory for highly simplified stimuli that can pro-
1338 vide relatively unbiased (compared with real-world experiences) measures of the relevant
1339 patterns and tendencies.

1340 The experiment we report in this paper was designed to help bridge some of this gap
1341 between naturalistic tasks and more traditional list-learning tasks. We had people study
1342 word lists similar to those used in classic memory studies, but we also systematically var-
1343 ied the lists' "richness" (by adding or removing visual features) and temporal structure
1344 (through order manipulations that varied over time and across experimental conditions).
1345 We found that participants' memory behaviors were sensitive to these manipulations.
1346 Some of the manipulations led to changes that were common across people (e.g., more
1347 temporal clustering when words' appearances were varied, enhanced memory for lists
1348 following an "event boundary," more feature clustering on order-manipulated lists, etc.).
1349 Other manipulations led to changes that were idiosyncratic (especially carryover effects
1350 from order manipulations; e.g., participants who remembered more words on early order-
1351 manipulated lists tended to show stronger feature clustering for their condition's feature
1352 dimension on late randomly ordered lists, etc.). We also found that participants remem-
1353 bered more words from lists that were sorted to align with their idiosyncratic clustering
1354 preferences. Taken together, our results suggest that our memories are susceptible to ex-
1355 ternal influences (i.e., to the statistical structure of ongoing experiences), but the effects of
1356 past experiences on future memory are largely idiosyncratic across people.

1357 **Potential applications**

1358 Every participant in our study encountered exactly the same words, split into exactly the
1359 same lists. But participants' memory performance, the orders in which they recalled the
1360 words, and the effects of early list manipulations on later lists all varied according to how
1361 we presented the to-be-remembered words.

1362 Our findings raise a number of exciting questions. For example, how far might these
1363 manipulations be extended? In other words, might there be more sophisticated or clever
1364 feature or order manipulations that one could implement to have stronger impacts on
1365 memory? Are there limits to how much impact (on memory performance and/or or-
1366 ganization) these sorts of manipulations can have? Are those limits universal across
1367 people, or are there individual differences (based on prior experiences, natural strate-
1368 gies, neuroanatomy, etc.) that impose person-specific limits on the potential impact of
1369 presentation-level manipulations on memory?

1370 Our findings indicate that the ways word lists are presented affects how people re-
1371 member them. To the extent that word list memory reflects memory processes that are
1372 relevant to real-world experiences, one could imagine potential real-world applications of
1373 our findings. For example, we found that participants remembered more words when the
1374 presentation order agreed with their memory fingerprints. If analogous fingerprints could
1375 be estimated for classroom content, perhaps they could be utilized manually by teachers,
1376 or even by automated content-presentation systems, to optimize how and what students
1377 remember.

1378 **Concluding remarks**

1379 Our work raises deep questions about the fundamental nature of human learning. What
1380 are the limits of our memory systems? How much does what we remember (and how we

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

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.
- Baddeley, A. D. (1968). Prior recall of newly learned items and the recency effect in free recall. *Canadian Journal of Psychology*, 22:157–163.
- 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.

- 1402 Beaman, C. P. and Jones, D. M. (1998). Irrelevant sound disrupts order information in
1403 free recall as in serial recall. *The Quarterly Journal of Experimental Psychology A: Human*
1404 *Experimental Psychology*, 51(3):615–636.
- 1405 Bonin, P., Thiebaut, G., Bugaiska, A., and Méot, A. (2022). Mixed evidence for a richness-of-
1406 encoding account of animacy effects in memory from the generation-of-ideas paradigm.
1407 *Current Psychology*, 41:1653–1662.
- 1408 Borges, M. A. and Mangler, G. (1972). Effect of within-category spacing on free recall.
1409 *Journal of Experimental Psychology*, 92(2):207–214.
- 1410 Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged
1411 associates. *Journal of General Psychology*, 49:229–240.
- 1412 Bousfield, W. A., Sedgewick, C. H., and Cohen, B. H. (1954). Certain temporal character-
1413 istics of the recall of verbal associates. *American Journal of Psychology*, 67:111–118.
- 1414 Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive*
1415 *Psychology*, 11(2):177–220.
- 1416 Bower, G. H., Lesgold, A. M., and Tieman, D. (1969). Grouping operations in free recall.
1417 *Journal of Verbal Learning and Verbal Behavior*, 8(4):481–493.
- 1418 Brigard, F. D. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*,
1419 3(420):1–3.
- 1420 Chiu, Y.-C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., and Sahakyan, L. (2021). Sepa-
1421 ration of item and context in item-method directed forgetting. *NeuroImage*, 235:117983.
- 1422 Chow, W.-Y., Momma, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory
1423 retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.

- 1424 Clayton, K. and Chattin, D. (1989). Spatial and semantic priming effects in tests of spa-
1425 tial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
1426 15(3):495–506.
- 1427 Clewett, D., DuBrow, S., and Davachi, L. (2019). Transcending time in the brain: how
1428 event memories are constructed from experience. *Hippocampus*, 29(3):162–183.
- 1429 Cofer, C. N., Bruce, D. R., and Reicher, G. M. (1966). Clustering in free recall as a function
1430 of certain methodological variations. *Journal of Experimental Psychology: General*, 71:858–
1431 866.
- 1432 D’Agostino, P. R. (1969). The blocked-random effect in recall and recognition. *Journal of*
1433 *Verbal Learning and Verbal Behavior*, 8:815–820.
- 1434 Dallett, K. M. (1964). Number of categories and category information in free recall. *Journal*
1435 *of Experimental Psychology*, 68:1–12.
- 1436 Darley, C. F. and Murdock, B. B. (1971). Effects of prior free recall testing on final recall
1437 and recognition. *Journal of Experimental Psychology: General*, 91:66–73.
- 1438 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct
1439 medial temporal lobe processes build item and source memories. *Proceedings of the*
1440 *National Academy of Sciences, USA*, 100(4):2157–2162.
- 1441 Donnelly, R. E. (1988). Priming effects in successive episodic tests. *Journal of Experimental*
1442 *Psychology: Learning, Memory, and Cognition*, 14:256–265.
- 1443 Drewnowski, A. and Murdock, B. B. (1980). The role of auditory features in memory span
1444 for words. *Journal of Experimental Psychology: Human Learning and Memory*, 6:319–332.

- 1445 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for
1446 the sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–
1447 1286.
- 1448 DuBrow, S. and Davachi, L. (2016). Temporal binding within and across events. *Neurobi-*
1449 *ology of Learning and Memory*, 134:107–114.
- 1450 DuBrow, S., Rouhani, N., Niv, Y., and Norman, K. A. (2017). Does mental context drift or
1451 shift? *Current Opinion in Behavioral Sciences*, 17:141–146.
- 1452 Eichenbaum, H. and Fortin, N. J. (2009). The neurobiology of memory based predictions.
1453 *Philosophical Transactions of the Royal Society of London Series B*, 364(1521):1183–1191.
- 1454 Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological*
1455 *Review*, 62:145–154.
- 1456 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory?
1457 *Psychological Science*, 22(2):243–252.
- 1458 Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall:
1459 a challenge for unitary models of recency. *Journal of Experimental Psychology: Learning,*
1460 *Memory, and Cognition*, 36:324–347.
- 1461 Farrell, S. (2014). Correcting the correction of conditional recency slopes. *Psychonomic*
1462 *Bulletin and Review*, 21:1174–1179.
- 1463 Fitzpatrick, P. C., Heusser, A. C., and Manning, J. R. (2023). Text embedding models yield
1464 high-resolution insights into conceptual knowledge from short multiple-choice quizzes.
1465 *PsyArXiv*, page doi.org/10.31234/osf.io/dh3q2.

- 1466 Flexser, A. J. and Tulving, E. (1982). Priming and recognition failure. *Journal of Verbal*
1467 *Learning and Verbal Behavior*, 21:237–248.
- 1468 Flores, S., Bailey, H. R., Eisenberg, M. L., and Zacks, J. M. (2017). Event segmentation
1469 improves event memory up to one month later. *Journal of Experimental Psychology:*
1470 *Learning, Memory, and Cognition*, 43(8):1183.
- 1471 Gershman, S. J., Schapiro, A. C., Hupbach, A., and Norman, K. A. (2013). Neural context
1472 reinstatement predicts memory misattribution. *The Journal of Neuroscience*, 33(20):8590–
1473 8595.
- 1474 Glenberg, A. M., Bradley, M. M., Kraus, T. A., and Renzaglia, G. J. (1983). Studies of the
1475 long-term recency effect: support for a contextually guided retrieval theory. *Journal of*
1476 *Experimental Psychology: Learning, Memory, and Cognition*, 12:413–418.
- 1477 Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
1478 prediction” task? individual variability in strategies for probabilistic category learning.
1479 *Learning and Memory*, 9:408–418.
- 1480 Gold, D. A., Zacks, J. M., and Flores, S. (2017). Effects of cues to event segmentation on
1481 subsequent memory. *Cognitive Research: Principles and Implications*, 2(1):1.
- 1482 Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder,
1483 A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto,
1484 C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A.,
1485 Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. (2021). Thinking
1486 ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*,
1487 page doi.org/10.1101/2020.12.02.403477.

- 1488 Gotts, S. J., Chow, C. C., and Martin, A. (2012). Repetition priming and repetition sup-
1489 pression: A case for enhanced efficiency through neural synchronization. *Cognitive*
1490 *Neuroscience*, 3(3-4):227–237.
- 1491 Griffiths, T. L. and Steyvers, M. (2003). Prediction and semantic association. *Advances in*
1492 *Neural Information Processing Systems*, 15.
- 1493 Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml,
1494 M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages
1495 2338–2342.
- 1496 Hargreaves, I. S., Pexman, P. M., Johnson, J. C., and Zdrazilova, L. (2012). Richer concepts
1497 are better remembered: number of features effects in free recall. *Frontiers in Human*
1498 *Neuroscience*, 6:doi.org/10.3389/fnhum.2012.00073.
- 1499 Healey, M. K. and Uitvlugt, M. G. (2019). The role of control processes in temporal and
1500 semantic contiguity. *Memory and Cognition*, 47:719–737.
- 1501 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail:
1502 a Python toolbox for analyzing and plotting free recall data. *Journal of Open Source*
1503 *Software*, 10.21105/joss.00424.
- 1504 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal
1505 behavioral and neural signatures of transforming experiences into memories. *Nature*
1506 *Human Behavior*, 5:905–919.
- 1507 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). HyperTools: a
1508 Python toolbox for gaining geometric insights into high-dimensional data. *Journal of*
1509 *Machine Learning Research*, 18(152):1–6.

- 1510 Hogan, R. M. (1975). Interitem encoding and directed search in free recall. *Memory and*
1511 *Cognition*, 3:197–209.
- 1512 Howard, M. W. and Kahana, M. J. (1999). Contextual variability and serial position effects
1513 in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:923–
1514 941.
- 1515 Howard, M. W. and Kahana, M. J. (2002a). A distributed representation of temporal
1516 context. *Journal of Mathematical Psychology*, 46:269–299.
- 1517 Howard, M. W. and Kahana, M. J. (2002b). When does semantic similarity help episodic
1518 retrieval? *Journal of Memory and Language*, 46:85–98.
- 1519 Huang, L., Holcombe, A. O., and Pashler, H. (2004). Repetition priming in visual search:
1520 episodic retrieval, not feature priming. *Memory and Cognition*, 32:12–20.
- 1521 Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental*
1522 *Psychology: General*, 137(2):324–347.
- 1523 Huber, D. E., Shiffrin, R. M., Lyle, K. B., and Ruys, K. I. (2001). Perception and preference
1524 in short-term word priming. *Psychological Review*, 108(1):149–182.
- 1525 Isarida, T. and Isarida, T. K. (2007). Environmental context effects of background color in
1526 free recall. *Memory and Cognition*, 35(7):1620–1629.
- 1527 Jenkins, J. J. and Russell, W. A. (1952). Associative clustering during recall. *Journal of*
1528 *Abnormal and Social Psychology*, 47:818–821.
- 1529 Jones, A. C. and Pyc, M. A. (2014). The production effect: costs and benefits in free recall.
1530 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):300–305.

- 1531 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing
1532 prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 1533 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*,
1534 24:103–109.
- 1535 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York,
1536 NY.
- 1537 Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychol-*
1538 *ogy*, 71:107–138.
- 1539 Kahana, M. J., Howard, M. W., and Polyn, S. M. (2008). Associative processes in episodic
1540 memory. In Roediger III, H. L., editor, *Cognitive Psychology of Memory*, pages 476–490.
1541 Elsevier, Oxford, UK.
- 1542 Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., and Yeshurun, Y. (2023). Deeper than
1543 you think: partisanship-dependent brain responses in early sensory and motor brain
1544 regions. *The Journal of Neuroscience*, pages doi.org/10.1523/JNEUROSCI.0895–22.2022.
- 1545 Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning
1546 of memories by context-based prediction error. *Proceedings of the National Academy of*
1547 *Sciences, USA*, In press.
- 1548 Kimball, D. R., Smith, T. A., and Kahana, M. J. (2007). The fSAM model of false recall.
1549 *Psychological Review*, 114(4):954–993.
- 1550 Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 1551 Lange, E. B. (2005). Disruption of attention by irrelevant stimuli in serial recall. *Journal of*
1552 *Memory and Language*, 43(4):513–531.

- 1553 Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases
1554 of human memory. *Current Opinion in Behavioral Sciences*, 32:111–119.
- 1555 Lohanas, L. J., Polyn, S. M., and Kahana, M. J. (2010). Modeling intralist and interlist effects
1556 in free recall. In *Psychonomic Society*, Saint Louis, MO.
- 1557 Luek, S. P., McLaughlin, J. P., and Cicala, G. A. (1971). Effects of blocking of input and
1558 blocking of retrieval cues on free recall learning. *Journal of Experimental Psychology*,
1559 91(1):159–161.
- 1560 Madan, C. R. (2021). Exploring word memorability: how well do different word properties
1561 explain item free-recall probability? *Psychonomic Bulletin and Review*, 28:583–595.
- 1562 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
1563 *Handbook of Human Memory*. Oxford University Press.
- 1564 Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
1565 function? *Psychological Review*, 128(4):711–725.
- 1566 Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. A.
1567 (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic*
1568 *Bulletin and Review*, 23(5):1534–1542.
- 1569 Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free
1570 recall. *Memory*, 20(5):511–517.
- 1571 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic
1572 memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.
- 1573 Manning, J. R., Notaro, G. M., Chen, E., and Fitzpatrick, P. C. (2022). Fitness tracking

1574 reveals task-specific associations between memory, mental health, and physical activity.
 1575 *Scientific Reports*, 12(13822):doi.org/10.1038/s41598-022-17781-0.

1576 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory pat-
 1577 terns in temporal lobe reveal context reinstatement during memory search. *Proceedings*
 1578 *of the National Academy of Sciences, USA*, 108(31):12893–12897.

1579 Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., and Kahana, M. J. (2012).
 1580 Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clus-
 1581 tering during memory search. *The Journal of Neuroscience*, 32(26):8871–8878.

1582 Marsh, J. E., Beaman, C. P., Hughes, R. W., and Jones, D. M. (2012). Inhibitory control in
 1583 memory: evidence for negative priming in free recall. *Journal of Experimental Psychology:*
 1584 *Learning, Memory, and Cognition*, 38(5):1377–1388.

1585 Marsh, J. E., Sörqvist, P., Hodgetts, H. M., Beaman, C. P., and Jones, D. M. (2015). Distraction
 1586 control processes in free recall: benefits and costs to performance. *Journal of Experimental*
 1587 *Psychology: Learning, Memory, and Cognition*, 41(1):118–133.

1588 Masicampo, E. J. and Sahakyan, L. (2014). Imagining another context during encoding off-
 1589 sets context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory,*
 1590 *and Cognition*, 40(6):1772–1777.

1591 Masís-Obando, R., Norman, K. A., and Baldassano, C. (2022). Scheme representations in
 1592 distinct brain networks support narrative memory during encoding and retrieval. *eLife*,
 1593 11:e70445.

1594 McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental*
 1595 *Psychology: Learning, Memory, and Cognition*, 20:507–520.

- 1596 Meinhardt, M. J., Bell, R., Buchner, A., and Röer, J. P. (2020). Adaptive memory: is
1597 the animacy effect on memory due to richness of encoding? *Journal of Experimental*
1598 *Psychology: Learning, Memory, and Cognition*, 46(3):416–426.
- 1599 Miller, J. F., Kahana, M. J., and Weidemann, C. T. (2012). Recall termination in free recall.
1600 *Memory and Cognition*, 40(4):540–550.
- 1601 Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman,
1602 S. J. (2017). The successor representation in human reinforcement learning. *Nature*
1603 *Human Behavior*, 1:680–692.
- 1604 Moran, R. and Goshen-Gottstein, Y. (2014). The conditional-recency dissociation is con-
1605 founded with nominal recency: should unitary models of memory still be devaluated?
1606 *Psychonomic Bulletin and Review*, 21:332–343.
- 1607 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental*
1608 *Psychology: General*, 64:482–488.
- 1609 Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy
1610 of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1611 Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhi-
1612 bitionless spreading activation and limited-capacity attention. *Journal of Experimental*
1613 *Psychology: General*, 106(3):226–254.
- 1614 Oberauer, K. and Lewandowsky, S. (2008). Forgetting in immediate serial recall: decay,
1615 temporal distinctiveness, or interference? *Psychological Review*, 115(3):544–576.
- 1616 Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., and Radvansky, G. A.
1617 (2016). Event boundaries and memory improvement. *Cognition*, 148:136–144.

- 1618 Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of
1619 context. *Trends in Cognitive Sciences*, 12:24–30.
- 1620 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). Task context and organization in
1621 free recall. *Neuropsychologia*, 47:2158–2163.
- 1622 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly*
1623 *Journal of Experimental Psychology*, 17:132–138.
- 1624 Puff, C. R. (1974). A consolidated theoretical view of stimulus-list organization effects in
1625 free recall. *Psychological Reports*, 34:275–288.
- 1626 Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of
1627 associative memory. In Bower, G. H., editor, *The Psychology of Learning and Motivation:*
1628 *Advances in Research and Theory*, volume 14, pages 207–262. Academic Press, New York,
1629 NY.
- 1630 Rabinowitz, J. C. (1986). Priming in episodic memory. *Journal of Gerontology*, 41:204–213.
- 1631 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:
1632 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.
- 1633 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition.
1634 *Current Opinion in Behavioral Sciences*, 17:133–140.
- 1635 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior.
1636 *Nature Reviews Neuroscience*, 13:713–726.
- 1637 Reinitz, M. T., Lammers, W. J., and Cochran, B. P. (1992). Memory-conjunction errors:
1638 miscombination of stored stimulus features can produce illusions of memory. *Memory*
1639 *and Cognition*, 20:1–11.

- 1640 Rissman, J., Eliassen, J. C., and Blumstein, S. E. (2003). An event-related fMRI investigation
1641 of implicit semantic priming. *Journal of Cognitive Neuroscience*, 15(8):1160–1175.
- 1642 Romney, A. K., Brewer, D. D., and Batchelder, W. H. (1993). Predicting clustering from
1643 semantic structure. *Psychological Science*, 4:28–34.
- 1644 Sahakyan, L. and Kelley, C. M. (2002). A contextual change account of the directed
1645 forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
1646 28(6):1064–1072.
- 1647 Sahakyan, L. and Smith, J. R. (2014). A long time ago, in a context far, far away: Retro-
1648 spective time estimates and internal context change. *Journal of Experimental Psychology:*
1649 *Learning, Memory, and Cognition*, 40(1):86–93.
- 1650 Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclo-*
1651 *pedic Reference*, 3:501–506.
- 1652 Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008). A context-based theory of
1653 recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.
- 1654 Sederberg, P. B., Miller, J. F., Howard, W. H., and Kahana, M. J. (2010). The tempo-
1655 ral contiguity effect predicts episodic memory performance. *Memory and Cognition*,
1656 38(6):689–699.
- 1657 Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of
1658 time. *Neural Computation*, 24:134–193.
- 1659 Shapiro, S. I. (1970). Isolation effects, free recall, and organization. *Journal of Psychology*,
1660 24:178–183.

- 1661 Shapiro, S. I. and Palermo, D. S. (1970). Conceptual organization and class membership:
1662 normative data for representatives of 100 categories. *Psychological Monograph Supple-*
1663 *ments*, 3(11):43.
- 1664 Sirotin, Y. B., Kimball, D. R., and Kahana, M. J. (2005). Going beyond a single list: modeling
1665 the effects of prior experience on episodic free recall. *Psychonomic Bulletin and Review*,
1666 12(5):787–805.
- 1667 Slamecka, N. J. and Barlow, W. (1979). The role of semantic and surface features in word
1668 repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 18:617–627.
- 1669 Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: a review and
1670 meta-analysis. *Psychonomic Bulletin and Review*, 8(2):203–220.
- 1671 Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009). A
1672 Bayesian analysis of dynamics in free recall. *Advances in Neural Information Processing*
1673 *Systems*, 22.
- 1674 Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., and Zacks, J. M. (2011).
1675 Changes in events alter how people remember recent information. *Journal of Cognitive*
1676 *Neuroscience*, 23(5):1052–1064.
- 1677 Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception
1678 affect memory encoding and updating. *Journal of Experimental Psychology: General*,
1679 138(2):236–257.
- 1680 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in*
1681 *Cognitive Sciences*, 22(3):201–212.
- 1682 Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *The*

- 1683 *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37:571–
1684 590.
- 1685 Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P.,
1686 and Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, 316(5821):76–
1687 82.
- 1688 Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press, New York, NY.
- 1689 Tulving, E. and Schacter, D. L. (1991). Priming and human memory systems. *Science*,
1690 247:301–305.
- 1691 Watkins, P. C., Mathews, A., Williamson, D. A., and Fuller, R. D. (1992). Mood-congruent
1692 memory in depression: emotional priming or elaboration? *Journal of Abnormal Psychol-*
1693 *ogy*, 101(3):581–586.
- 1694 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American*
1695 *Journal of Psychology*, 35:396–401.
- 1696 Whitely, P. L. (1927). The dependence of learning and recall upon prior intellectual activi-
1697 ties. *Journal of Experimental Psychology: General*, 10:489–508.
- 1698 Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming.
1699 *Current Opinion in Neurobiology*, 8(2):227–233.
- 1700 Xu, X., Zhu, Z., and Manning, J. R. (2023). The psychological arrow of time drives
1701 temporal asymmetries in retrodicting versus predicting narrative events. *PsyArXiv*,
1702 page doi.org/10.31234/osf.io/yp2qu.
- 1703 Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., and Hasson, U.

1704 (2017). Same story, different story: the neural representation of interpretive frameworks.
1705 *Psychological Science*, 28(3):307–319.

1706 Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., and
1707 Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal*
1708 *of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792–804.

1709 Zhang, Q., Griffiths, T. L., and Norman, K. A. (2023). Optimal policies for free recall.
1710 *Psychological Review*, 130(4):1104–1125.

1711 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018).
1712 Is automatic speech-to-text transcription ready for use in psychological experiments?
1713 *Behavior Research Methods*, 50:2597–2605.

1714 Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation
1715 models in narrative comprehension: an event-indexing model. *Psychological Science*,
1716 6(5):292–297.

1717 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension
1718 and memory. *Psychological Bulletin*, 123(2):162–185.