

Carryover effects in free recall reveal how prior experiences influence memories of new experiences

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

¹Dartmouth College

²Princeton University

³Akili Interactive

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

Abstract

We perceive, interpret, and remember ongoing experiences through the lens of our prior experiences. Inferring that we are 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 of 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 the orders in which the words were studied. We found that these order manipulations affected not only how the participants recalled the ordered lists, but also how they recalled later randomly ordered lists. Our work shows how structure in our ongoing experiences can exert influence on how we remember unrelated subsequent experiences.

17 Introduction

18 Experience is subjective: different people who encounter identical physical experiences
19 can take away very different meanings and memories. One reason is that our subjective ex-
20 periences in the moment are shaped in part the idiosyncratic prior experiences, memories,
21 goals, thoughts, expectations, and emotions that we bring with us into the present moment.
22 These factors collectively define a *context* for our experiences[?]. situation models: forming
23 expectations, predicting ambiguous future experiences The contexts we encounter help
24 us to construct *situation models*^{??} or *schemas*^{??} that describe how experiences are likely to
25 unfold based on our prior experiences with similar contextual cues. For example, when
26 we enter a sit-down restaurant, we might expect to be seated at a table, given a menu,
27 and served food. Priming someone to expect a particular situation or context can also
28 influence how they resolve potential ambiguities in their ongoing experiences, including
29 ambiguous movies and narratives[?].

30 Our understanding of how we form situation models and schemas, and how they in-
31 teract with our subjective experiences and memories, is constrained in part by substantial
32 differences in how we study these processes. Situation models and schemas are most often
33 studied using “naturalistic” stimuli such as narratives and movies^{???}. In contrast, our
34 understanding of how we organize our memories has been most widely studied using
35 more traditional paradigms like free recall of random word lists[?]. In free recall, partici-
36 pants study lists of items and are instructed to recall the items in any order they choose.
37 The orders in which words come to mind can provide insights into how participants have
38 organized their memories of the studied words. Because random word lists are unstruc-
39 tured by design, it is not clear if or how non-trivial situation models might apply to these
40 stimuli. Nevertheless, there are *some* commonalities between memory for word lists and
41 memory for real-world experiences.

42 Like remembering real-world experiences, remembering words on a studied list re-
43 quires distinguishing the current list from the rest of one's experience. To model this
44 fundamental memory capability, cognitive scientists have posited the existence of a spe-
45 cial representation, called *context*, that is associated with each list. According to early
46 theories e.g.^{??} context representations are composed of many features which fluctuate
47 from moment to moment, slowly drifting through a multidimensional feature space. Dur-
48 ing recall, this representation forms part of the retrieval cue, enabling us to distinguish
49 list items from non-list items. Understanding the role of context in memory processes is
50 particularly important in self-cued memory tasks, such as *free recall*, where the retrieval
51 cue is "context" itself.

52 Over the past half-century, context-based models have enjoyed impressive success at
53 explaining many stereotyped behaviors observed during free recall and other list-learning
54 tasks^{????????}. These phenomena include the well-known recency and primacy
55 effects (superior recall of items from the end and, to a lesser extent, from the beginning of
56 the study list), as well as semantic and temporal clustering effects[?]. The contiguity effect
57 is an example of temporal clustering, which is perhaps the dominant form of organization
58 in free recall. This effect can be seen in the tendency for people to successively recall items
59 that occupied neighboring positions in the study list. For example, if a list contained the
60 sub-sequence "ABSENCE HOLLOW PUPIL" and the participant recalls the word "HOLLOW", it is
61 far more likely that the next response will be either "PUPIL" or "ABSENCE" than some other
62 list item[?]. In addition, there is a strong forward bias in the contiguity effect: subjects
63 make forward transitions (i.e., "HOLLOW" followed by "PUPIL") about twice as often as
64 they make backward transitions, despite an overall tendency to begin recall at the end of
65 the list. There are also striking effects of semantic clustering^{????}, whereby the recall
66 of a given item is more likely to be followed by recall of a similar or related item than

67 a dissimilar or unrelated one. In general, people organize memories for words along a
68 wide variety of stimulus dimensions. As captured by models like the *Context Maintenance*
69 *and Retrieval Model*[?], the stimulus features associated with each word (e.g. the word's
70 meaning, font size, font color, location on the screen, size of the object the word represents,
71 etc.) are incorporated into the participant's mental context representation^{????}. During
72 a memory test, any of these features may serve as a memory cue, which in turn leads the
73 participant to recall in succession words that share stimulus features.

74 A key mystery is whether the sorts of situation models and schemas that people use to
75 organize their memories of real-world experiences might map onto the clustering effects
76 that reflect how people organize their memories for word lists. On one hand, situation
77 models and clustering effects both reflect statistical regularities in ongoing experience.
78 Our memory systems exploit these regularities when generating inferences about the
79 unobserved past and yet-to-be-experienced future^{????}. On the other hand, the rich
80 structure of real-world experiences and other naturalistic stimuli that enable people to
81 form deep and meaningful situation models and schemas have no obvious analog in
82 simple word lists. Often lists in free recall studies are explicitly *designed* to be devoid of
83 exploitable temporal structure, for example by sorting the words in a random order[?].

84 We designed an experimental paradigm to explore how people organize their mem-
85 ories for simple stimuli (word lists) whose temporal properties change across different
86 "situations," analogous to how the content of real-world experiences change across dif-
87 ferent real-world situations. We asked participants to study and freely recall a series
88 of word lists (Fig. ??). Across the different conditions in the experiment, we varied the
89 lists' presentation orders in different ways across lists. The studied items (words) were
90 designed to vary along three general dimensions: semantic (word *category*, and physical
91 *size* of the referent), lexicographic (word *length* and *first letter*), and visual (font *color* and

the onscreen *location* of each word). In our main manipulation conditions, we asked participants to study and recall eight lists whose items were sorted by a target feature (e.g., word category). Next, we asked them to study and recall an additional eight lists whose items had the same features, but that were sorted in a random temporal order. We were interested in how these order manipulations affected participants' recall behaviors on early (sorted) lists, as well as how order manipulations on early lists affected recall behaviors on later (unsorted) lists. We used a series of control conditions as a baseline; in these control conditions all of the lists were sorted randomly, but we manipulated the presence or absence of the visual features. Finally, in an *adaptive* experimental condition we used participants' recall behaviors on early lists to manipulate, in real-time, the presentation orders of subsequent lists. In this adaptive condition, we sought to identify potential commonalities within and across participants in how people organized their memories and how those organizational tendencies affect overall performance.

Materials and methods

Participants

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

Participants received course credit for enrolling in our study. We asked each participant to fill out a demographic survey that included information about their self-reported age, gender, ethnicity, race, education, vision, reading impairments, medications or recent

injuries, coffee consumption on the day of testing, and level of alertness at the time of testing. All components of the demographics survey were optional. One participant elected not to fill out any part of the demographic survey, and all other participants report some or all of their requested demographic information.

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

Participants were assigned to experimental conditions based loosely on their date of participation. (This aspect of our procedure helped us to more easily synchronize the experiment databases across multiple testing computers.) Of the 490 participants who opted to fill out the demographics survey, reported ages ranged from 17 to 31 years (mean: 19.1; standard deviation: 1.356). A total of 318 participants reported their gender as female, 170 as male, and 2 participants declined to report their gender. A total of 442 participants reported their ethnicity as “not Hispanic or Latino,” 39 as “Hispanic or Latino,” and 9 declined to report their ethnicity. Participants reported their races as White (345 participants), Asian (120 participants), Black or African American (31 participants), American Indian or Alaska Native (11 participants), Native Hawaiian or Other Pacific Islander (4 participants), Mixed race (3 participants), Middle Eastern (1 participant), and Arab (1

140 participant). A total of 5 participants declined to report their race. We note that several
141 participants reported more than one of racial category. Participants reported their high-
142 est degrees achieved as “Some college” (359 participants), “High school graduate” (117
143 participants), “College graduate” (7 participants), “Some high school” (5 participants),
144 “Doctorate” (1 participant), and “Master’s degree” (1 participant). A total of 482 partici-
145 pants reported no reading impairments, and 8 reported mild reading impairments such
146 as mild dyslexia. A total of 489 participants reported having normal color vision and 1
147 participant reported that they were color blind. A total of 482 participants reported taking
148 no prescription medications and having no recent injuries; 4 participants reported having
149 ADHD, 1 reported having dyslexia, 1 reported having allergies, 1 reported a recently
150 torn ACL/MCL, and 1 reported a concussion from several months prior. The participants
151 reported consuming 0 – 3 cups of coffee prior to the testing session (mean: 0.32 cups;
152 standard deviation: 0.58 cups). Participants reported their current level of alertness, and
153 we converted their responses to numerical scores as follows: “very sluggish” (-2), “a little
154 sluggish” (-1), “neutral” (0), “a little alert” (1), and “very alert” (2). Across all partici-
155 pants, the full range of alertness levels were reported (range: -2 – 2; mean: 0.35; standard
156 deviation: 0.89).

157 We dropped from our dataset the 1 participant who reported abnormal color vision, as
158 well as 39 participants whose data were corrupted due to technical failures while running
159 the experiment or during the daily database merges. In total, this left usable data from
160 452 participants, broken down by experimental condition as follows: feature rich (67
161 participants), reduced (61 participants), reduced (late) (41 participants), reduced (early),
162 (42 participants), category (30 participants), size (30 participants), length (30 participants),
163 first letter (30 participants), color (31 participants), location (30 participants), and adaptive
164 (60 participants). The participant who declined to fill out their demographic survey

165 participated in the location condition, and we verified verbally that they had normal color
166 vision.

167 **Experimental design**

168 Our experiment is a variant of the classic free recall paradigm that we term *feature-rich free*
169 *recall*. In feature-rich free recall, participants study 16 lists, each comprised of 16 words that
170 vary along a number of stimulus dimensions (Fig. ??). The stimulus dimensions include
171 two semantic features related to the *meanings* of the words (semantic category, referent
172 object size), two lexicographic features related to the *letters* that make up the words (word
173 length in number of letters, identity of the word's first letter), and two visual features
174 that are independent of the words themselves (text color, presentation location). Each
175 list contains four words from each of four different semantic categories and two object
176 sizes; all other stimulus features are randomized. After studying each list, the participant
177 attempts to recall as many words as they can from that list, in any order they choose.
178 Because each individual word is associated with several well-defined (and quantifiable)
179 features, and because each list incorporates a diverse mix of feature values along each
180 dimension, this allows us to evaluate participants' memory fingerprints in rich detail.

181 **Stimuli**

182 Stimuli in our paradigm were 256 English words selected in a previous study⁷. The words
183 all referred to concrete nouns, and were chosen from 15 unique semantic categories: body
184 parts, building-related, cities, clothing, countries, flowers, fruits, insects, instruments,
185 kitchen-related, mammals, (US) states, tools, trees, and vegetables. We also tagged each
186 word according to the approximate size of the object the word referred to. Words were
187 labeled as "small" if the corresponding object was likely able to "fit in a standard shoebox"



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

188 or “large” if the object was larger than a shoebox. Semantic categories varied in how many
189 object sizes they reflected (mean number of different sizes per category: 1.33; standard
190 deviation: 0.49). The numbers of words in each semantic category also varied from 12
191 – 28 (mean number of words per category: 17.07; standard deviation number of words:
192 4.65). We also identified lexicographic features for each word, including the words’ first
193 letters and lengths (i.e., number of letters). Across all categories, all possible first letters
194 were represented except for ‘Q’ (average number of unique first letters per category: 11;
195 standard deviation: 2 letters). Word lengths ranged from 3 – 12 letters (average: 6.17
196 letters; standard deviation: 2.06 letters).

197 We assigned the categorized words into a total of 16 lists with several constraints.
198 First, we required that each list contained words from exactly 4 unique categories, each
199 with exactly 4 exemplars from each category. Second, we required that (across all words
200 on the list) at least one instance of both object sizes were represented. On average, each
201 category was represented in 4.27 lists (standard deviation: 1.16 lists). Aside from these
202 two constraints, we assigned each word to a unique list. After random assignment, each
203 list contained words with an average of 11.13 unique starting letters (standard deviation:
204 1.15 letters) and an average word length of 6.17 letters (standard deviation: 0.34 letters).

205 The above assignments of words to lists was performed once across all participants,
206 such that every participant studied the same set of 16 lists. In every condition we random-
207 ized the study order of these lists across participants. For participants in some conditions,
208 on some lists, we also randomly varied two additional visual features to each word: the
209 presentation font color, and the word’s onscreen location. These attributes were assigned
210 independently for word (and for every participant) at the times the words were displayed
211 onscreen. These visual features were varied for words in all lists and conditions except for
212 the “reduced” condition (all lists), the first eight lists of the “reduced (early)” condition,

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

215 To assign a random font color to each word, we selected three integers uniformly
216 and at random between 0 and 255, corresponding to the red (r), green (g), and blue (b)
217 color channels for that word. To assign random presentation locations to each word, we
218 selected two floating point numbers uniformly at random (one for the word’s horizontal
219 x coordinate and the other for its vertical y coordinate). The bounds of these coordinates
220 were selected to cover the entire visible area of the display without cutting off any part of
221 the words. The words were shown on 27 in (diagonal) Retina 5K iMac displays (resolution:
222 5120×2880 pixels).

223 Most of the experimental manipulations we carried out entailed presenting or sorting
224 the presented words differently on the first eight lists participants studied (which we call
225 *early* lists) versus on the final eight lists they studied (*late* lists). Since every participant
226 studied exactly 16 lists, using this terminology every list was either “early” or “late”
227 depending on its order in the list study sequence.

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

229 Our experimental paradigm incorporates the Google Cloud Speech API speech-to-text
230 engine⁷ to automatically transcribe participants’ verbal recalls into text. This allows
231 recalls to be transcribed in real time— a distinguishing feature of the experiment; in typical
232 verbal recall experiments the audio data must be parsed manually. In prior work, we
233 used a similar experimental setup (equivalent to the “reduced” condition in the present
234 study) to verify that the automatically transcribed recalls were sufficiently close to human-
235 transcribed recalls to yield reliable data⁷. This real-time speech processing component of
236 the paradigm plays an important role in the “adaptive” condition of the experiment, as

237 described below.

238 **Random conditions (Fig. ??, top four rows)**

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

248 In the *reduced (early)* condition, we followed the “reduced” procedure (presenting each
249 word in black at the center of the screen) for early lists, and followed the “feature rich”
250 procedure (presenting each word in a random color and location) for late lists. Finally, in
251 the *reduced (late)* condition, we followed the feature rich procedure for earlylists and the
252 reduced procedure for late lists.

253 **Order manipulation conditions (Fig. ??, middle six rows)**

254 Each of six *order manipulation* conditions used a different feature-based sorting procedure
255 to order words on early lists, where each sorting procedure relied on one relevant feature
256 dimension. All of the irrelevant features varied freely across words on early lists, in
257 that we did not consider irrelevant features in ordering the early lists. However, some
258 features were correlated— for example, some semantic categories of words referred to
259 objects that tended to be a particular size, which means that category and size are not

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

263 **Defining feature-based distances.** Sorting words according to a given relevant feature
264 requires first defining a distance function for quantifying the dissimilarity between each
265 pair of features. This function varied according to the type of features. Semantic features
266 (category and size) are *categorical*. For these features, we defined a binary distance function:
267 two words were considered to “match” (i.e., have a distance of 0) if their labels are the
268 same (i.e., both from the same semantic category or both of the same size). If two words'
269 labels were different for a given feature, we defined the words to have a distance of 1 for
270 that feature. Lexicographic features (length and first letter) are *discrete*. For these features
271 we defined a discrete distance function. Specifically, we defined the distance between
272 two words as either the absolute difference between their lengths, or the absolute distance
273 between their starting letters in the English alphabet, respectively. For example, two
274 words that started with the same letter would have a “first letter” distance of 0, and words
275 starting with ‘J’ and ‘A’ would have a first letter distance of 9. Because words' lengths
276 and letters' positions in the alphabet are always integers, these discrete distances always
277 take on integer values. Finally, the visual features (color and location) are *continuous* and
278 *multivariate*, in that each “feature” takes on multiple (positive) real values. We defined the
279 “color” and “location” distances between two words as the Euclidean distances between
280 their (r, g, b) color or (x, y) location vectors, respectively. Therefore the color and location
281 distance measures always take on positive real values (upper bounded at 441.67 for color, or
282 27 in for location, reflecting the distances between the corresponding maximally different
283 vectors).

284 **Constructing feature-sorted lists.** Given a list of words, a relevant feature, and each
 285 word’s value(s) for that feature, we developed a stochastic algorithm for (noisily) sorting
 286 the words. First, we choose a word uniformly at random from the set of candidates. Next,
 287 we compute the distances between the chosen word’s feature(s) and the corresponding
 288 feature(s) of all yet-to-be-presented words. Third, we convert these distances (between the
 289 previously presented word’s feature values, a , and the candidate word’s feature values, b)
 290 to similarity scores:

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

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

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

295 where in the demoniator, i takes on each of the n feature values of the to-be-presented
 296 words. The resulting set of normalized similarity scores sums to one.

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

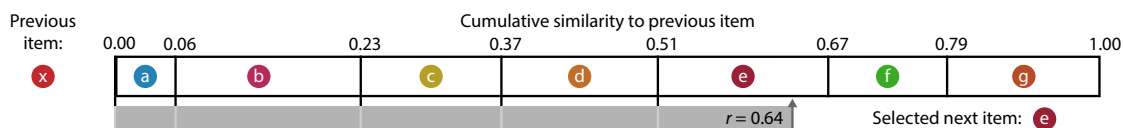


Figure 2: Generating stochastic feature-sorted lists. For a given feature dimension (e.g., color), we compute the similarity (Eqn. ??) between the feature value(s) of the previous item, x , and all yet-to-be-presented items ($a - g$). Next, we normalize these similarity scores so that they sum to one. We lay in sequence a set of “sticks,” one for each candidate item, whose lengths are equal to these normalized similarity scores. Note that the combined lengths of these sticks is one. To select the next to-be-presented item, we draw a random number, r , from the uniform distribution bounded between 0 and 1 (inclusive). The identity of the next item is given by the stick adjacent to an indicator that moves distance r (starting from 0) along the sequence of sticks. In this case, the next to-be-presented item is e . Note that each item’s chances of selection is proportional to its similarity to the previous item, along the given feature dimension.

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

308 Adaptive condition

309 We designed the *adaptive* experimental condition to study the effect on memory for infor-
 310 mation that matched (or mismatched) the ways participants “naturally” organized their
 311 memories of the lists they studied. Like the other conditions, all participants in the adap-
 312 tive condition studied a total of 16 lists, in a randomized order. We varied the words’ colors
 313 and locations for every word presentation, as in the feature rich and order manipulation
 314 conditions.

315 All participants in the adaptive condition began the experiment by studying a set of
 316 four *initialization* lists. Words and features on these lists were presented in a randomized
 317 order (computed independently for each participant). These initialization lists were used
 318 to estimate each participant’s “memory fingerprint,” defined below. At a high level, a
 319 participant’s memory fingerprint describes how they prioritize different semantic, lexico-

320 graphic, and/or visual features when they organize their memories.

321 Next, participants studied a sequence of 12 lists in three batches of 4 lists each. These
322 batches came in three types: *random*, *stabilize*, and *destabilize*. The batch types determined
323 how words on the lists in that batch were ordered. Lists in each batch were always
324 presented consecutively (e.g., a participant might receive four random lists, followed
325 by four stabilize lists, followed by four destabilize lists). The batch orders were evenly
326 counterbalanced across participants: there are six possible orderings of the three batches,
327 and 10 participants were randomly assigned to each ordering sub-condition.

328 Lists in the random batches were sorted randomly (as on the initialization lists and in
329 the feature rich condition). Lists in the stabilize and destabilize batches were sorted in
330 ways that either matched or mismatched each participant’s memory fingerprint, respec-
331 tively. Our procedures for computing participants’ memory fingerprints and ordering the
332 stabilize and destabilize lists are described next.

333 **Feature clustering scores (uncorrected).** Feature clustering describes a participant’s ten-
334 dency to recall similar presented items together in their recall sequences, where “simi-
335 larity” considers one given feature dimension (e.g., category, color, etc.). Computing the
336 clustering score for one feature dimension starts by considering those feature values from
337 the first word the participant recalled on the list. Next, we sort all not-yet-recalled words
338 in ascending order according to their feature-based distance to the just-recalled item (see
339 *Defining feature-based distances*). We then compute the percentile rank of the observed next
340 recall. We averaged these percentile ranks across all of the participant’s recalls for the
341 current list to obtain a single uncorrected clustering score for the list, for the given feature
342 dimension. We repeat this process for each feature dimension in turn to obtain a single
343 uncorrected clustering score for each list, for each feature dimension.

344 **Permutation-corrected feature clustering scores.**

345 **Memory fingerprints.**

346 **Online “fingerprint” analysis.**

347 **Ordering “stabilize” lists by an estimated fingerprint.**

348 **Ordering “destabilize” lists by an estimated fingerprint.**

349 **Analyses**

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

351 Probability of first recall curves^{???} reflect the probability that an item will be recalled
352 first, as a function of its serial position during encoding. To carry out this analysis, we
353 initialized (for each participant) a number-of-lists (16) by number-of-words-per-list (16)
354 matrix of zeros. Then, for each list, we found the index of the word that was recalled first,
355 and we filled in that position in the matrix with a 1. Finally, we averaged over the rows
356 of the matrix to obtain a 1 by 16 array of probabilities, for each participant. We used an
357 analogous procedure to compute probability of n^{th} recall curves for each participant.
358 Specifically, we filled in the corresponding matrices according to the n^{th} recall on each
359 list that each participant made. When a given participant had made fewer than n recalls
360 for a given list, we simply excluded that list from our analysis when computing that
361 participant’s curve(s).

362 **Lag-conditional response probability curve**

363 The lag-conditional probability (lag-CRP) curve⁷ reflects the probability of recalling a
364 given item after the just-recalled item, as a function of their relative encoding positions
365 (lag). In other words, a lag of 1 indicates that a recalled item was presented immediately
366 after the previously recalled item, and a lag of 3 indicates that a recalled item came 3
367 items before the previously recalled item. For each recall transition (following the first
368 recall), we computed the lag between the just-recalled word's presentation position and the
369 next-recalled word's presentation position. We computed the proportions of transitions
370 (between successively recalled words) for each lag, normalizing for the total numbers of
371 possible transitions. In carrying out this analysis, we excluded all incorrect recalls and
372 successive repetitions (e.g., recalling the same word twice in a row). This yielded, for
373 each list, a 1 by number-of-lags (−15 to +15; 30 lags in total, excluding lags of 0) array of
374 conditional probabilities. We averaged these probabilities across lists to obtain a single
375 lag-CRP for each participant.

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

386 our permutation correction procedure to obtain a corrected clustering score.

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

388 **Serial position curve**

389 Serial position curves[?] reflect the proportion of participants who remember each item as a
390 function of the item's serial position during encoding. For each participant, we initialized
391 a number-of-lists (16) by number-of-words-per-list (16) matrix of zeros. Then, for each
392 correct recall, we identified the presentation position of the word and entered a 1 into that
393 position (row: list; column: presentation position) in the matrix. This resulted in a matrix
394 whose entries indicated whether or not the words presented at each position, on each list,
395 were recalled by the participant (depending on whether the corresponding entries were
396 set to one or zero). Finally, we averaged over the rows of the matrix to yield a 1 by 16 array
397 representing the proportion of words at each position that the participant remembered.

398 **Identifying event boundaries**

399 **Results**

400 Figure S3.

401 Figure S7.

402 Figure S4.

403 **Discussion**

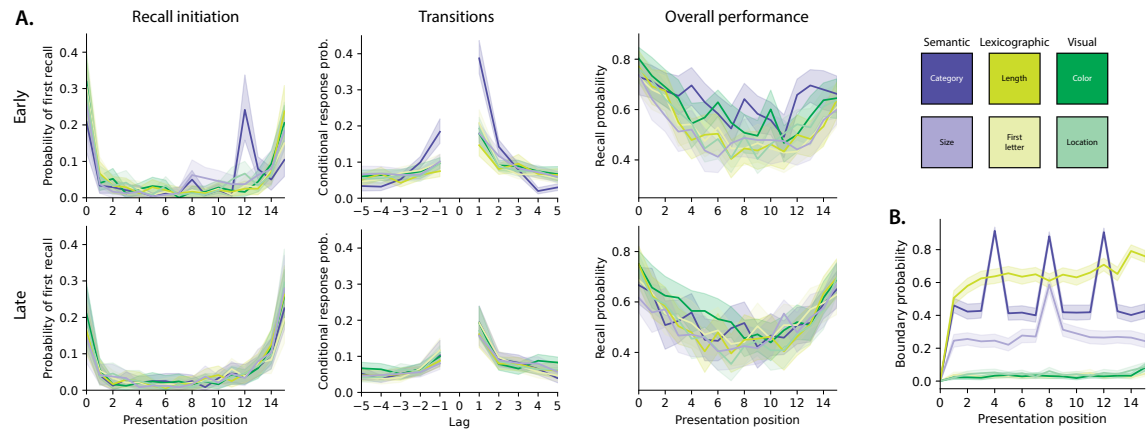


Figure 3: Recall dynamics in feature rich free recall (order manipulation conditions). **A.** Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling each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

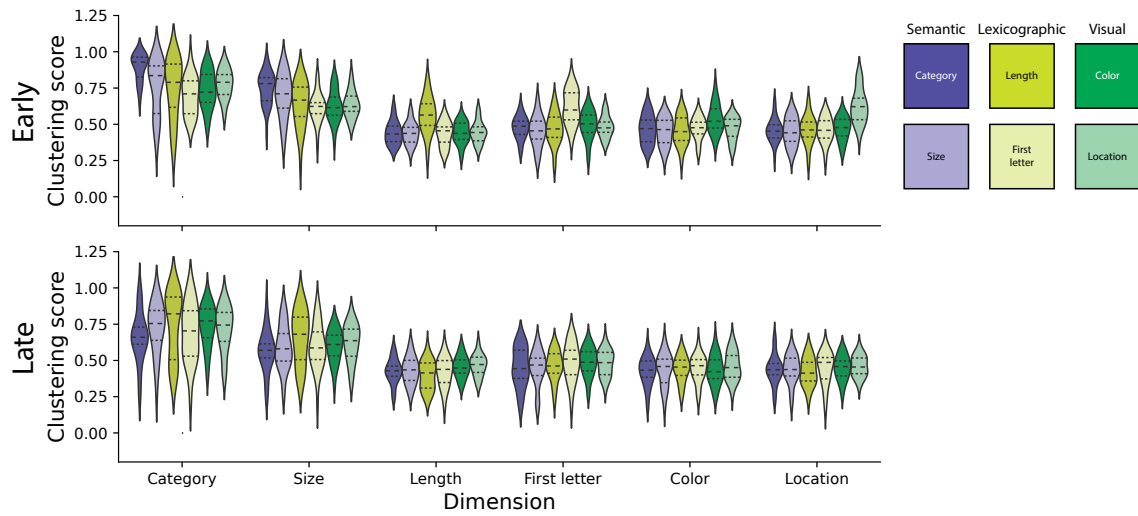


Figure 4: Memory “fingerprints” (order manipulation conditions). The across-participant distributions of clustering scores for each feature type (x -coordinate) are displayed for each experimental condition (color), separately for order manipulation (early, top) and randomly ordered (late, bottom) lists.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

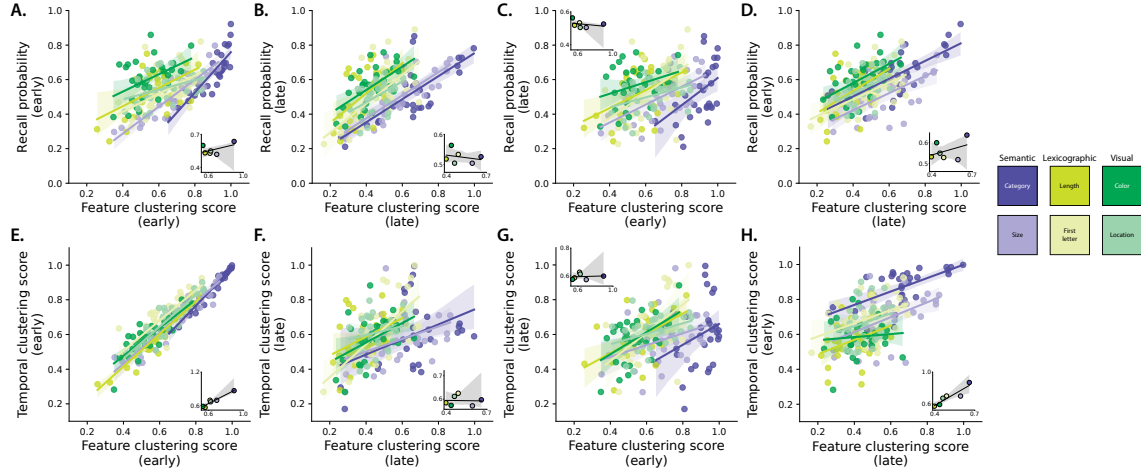


Figure 5: Interactions between feature clustering, recall probability, and contiguity. A. Recall probability versus feature clustering scores for order manipulation (early) lists. B. Recall probability versus feature clustering for randomly ordered (late) lists. C. Recall probability on late lists versus 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.

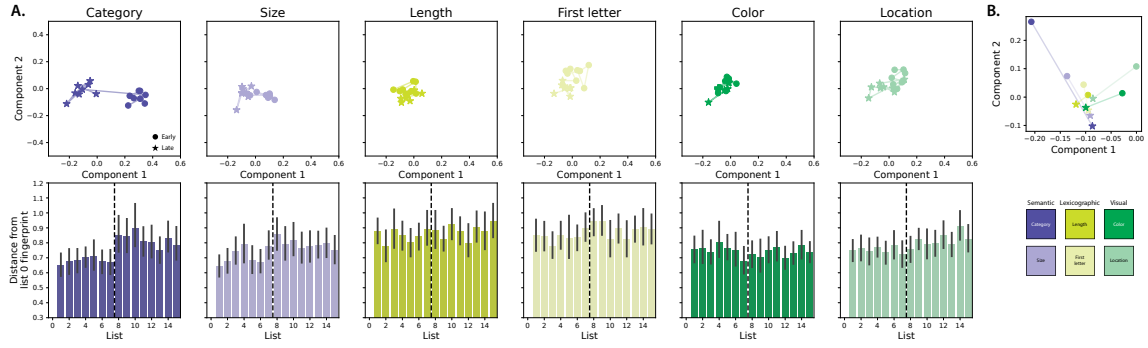


Figure 6: Memory fingerprint dynamics (order manipulation conditions). **A.** Each column (and color) reflects an experimental condition. In the top panels, each marker displays a 2D projection of the (across-participant) average memory fingerprint for one list. Order manipulation (early) lists are denoted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

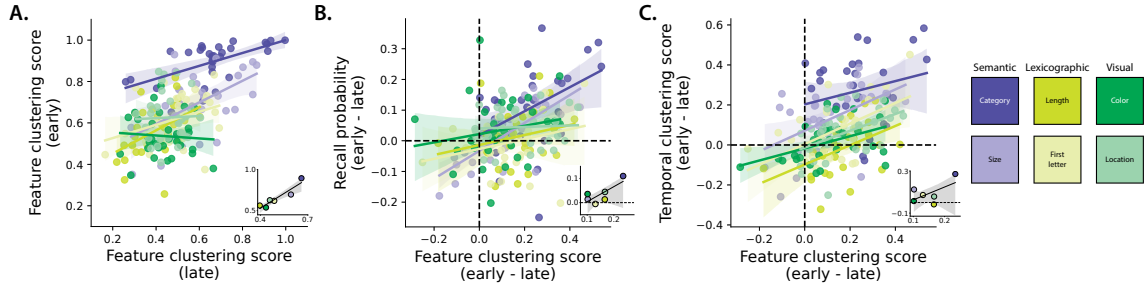


Figure 7: Feature clustering carryover effects. **A.** Feature clustering scores for order manipulation (early) versus randomly ordered (late) lists. **B.** Accuracy differences (on early versus late lists) versus feature clustering “carryover” (defined as the differences between the average clustering 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.

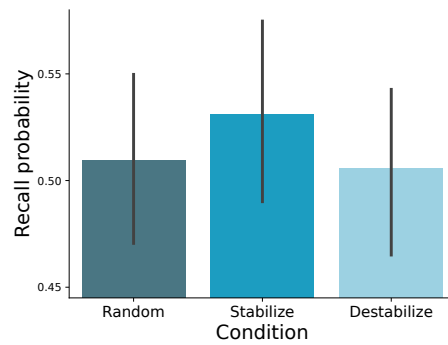


Figure 8: Recall performance (adaptive conditions). The bars display the average probability of recall (taken across words, lists, and participants) for lists from each adaptive condition. Error bars denote bootstrap-estimated 95% confidence intervals. For additional details about participants' behavior and performance during the adaptive conditions, see Figure S2.