Organizational Dynamics of Memory Across Days

Daniella Rafla, Brandon S. Katerman, David Halpern, and Michael J. Kahana

Department of Psychology, University of Pennsylvania

Author Note

The authors gratefully acknowledge support from National Institutes of Health (grant MH55687). Correspondence concerning this article may be addressed to Michael J. Kahana. E-mail: kahana@psych.upenn.edu

Abstract

When individuals repeatedly study and recall information across multiple learning trials their responses exhibit increasing levels of subjective organization. Whereas classic studies investigated the evolution of organization across lists within the short-time span of a single session, here we ask how memory changes over many days. Specifically, we examine how semantic, temporal, and subjective organization during a recall period shapes memory after days of intervening cognitive activity. Analyzing data from two multi-session free recall experiments, we find that subjects demonstrate a strong tendency to cluster recalls based on previous output order, with this effect strengthening across sessions. In line with the idea that thoughts become memories, we show that even false memories produced on a given session tend to re-occur on subsequent days. Our results attest to the crucial role that retrieval plays in shaping long-term episodic memory.

Organizational Dynamics of Memory Across Days

Introduction

Much of our knowledge concerning human memory derives from list memory paradigms, such as free recall, cued recall, and item recognition (Murdock, 1974). Each of these paradigms typically involves a study phase, during which a subject experiences a series of items, followed by a test phase, in which the experimenter assesses the subject's memory for the previously studied items. Each item within the list constitutes an event within the larger episode of the target list, and researchers use the term episodic memory to refer to our ability to uniquely retrieve these events (Tulving, 1983). In his influential systems analysis of memory, Tulving (1985) distinguished episodic memory from other forms of memory, including semantic memory and perceptual priming.

Here we investigate the dynamics of memory retrieval in a setting that aims to bridge the episodic memory literature with studies that seek to explore learning on much longer time scales. Specifically, we ask how people's memory for random word lists evolves over multiple days as a research subject gains greater and greater knowledge of the pool of potential memoranda. To do this, we repeat the same multi-list episodic memory paradigm over multiple days, with each session employing an identical pool of items. Once each session, we interrogate memory for the entire pool of words.

Students of memory have long known that semantic knowledge can exert a powerful influence on episodic recall (e.g., W. A. Bousfield, Sedgewick, & Cohen, 1954). Yet, over a century of scientific toil has produced scant knowledge of whether or how a subject's experience studying or recalling a list of words shapes their memory for those words after days of intervening cognitive activity. We examine such long-term effects on episodic memory in the setting of free recall, where subjects attempt to recall a large list of previously experienced items in any order they wish (Kirkpatrick, 1894). We hypothesized that the order in which words come to mind on day 1 significantly impacts the order in which the same words come to mind on day 2, even when recall on day 2 occurs without

reference to day 1.

Although a vast literature has characterized the accumulation of knowledge across repeated learning trials, most learning paradigms either involve intentional learning (e.g., Klein, Addis, & Kahana, 2005; Siegel & Kahana, 2014), or least involve consistent mapping across trials, as in the case of the Hebb (1961) repetition effect. However, information encoded for an immediate retention test does not vanish immediately following a recall test. Indeed, data from several well-known paradigms attest to the durability of memory following a single immediate recall trial. For example, in Craik's (1970) final free recall procedure, subjects undergo a series of immediate recall trials of (typically) random word lists. Then, at the end of the entire session, the experimenter asks subjects to freely recall all of the items from the previously experienced lists. For those items successfully recalled on their original lists, performance on final free recall is fairly good and shows a marked recency effect across lists (Kuhn, Lohnas, & Kahana, 2018). Data on prior-list intrusions reveals that memories of previously studied items persist across many lists. Such intrusions occur with decreasing probability as the number of intervening lists increases (Zaromb et al., 2006; Healey & Kahana, 2014). Similar effects appear in studies of short-term item recognition, where subjects tend to endorse lures when they occurred as a target on a recent prior list (Monsell, 1978).

Although many prior studies have documented the transfer of learning across trials, little is known about longer-term transfer in short-term memory paradigms. Here we reanalyze final-free recall data from two unique datasets that allowed us to evaluate whether the order in which subjects recalled items on prior lists shapes their memory for those same items on subsequent days, and under conditions that lack any explicit direction to encode items for long-term retention or to direct retrieval towards information learned on prior sessions. Both experiments contained a 10-minute free recall task in which subjects attempted to recall as many items as they could from all prior studied lists.

Methods

We first analyzed data from the unconventional multi-session free recall experiment reported in (Katerman, Li, Pazdera, Keane, & Kahana, 2022). This experiment featured a study-test delay of a day or more. We examined the dynamics of memory organization across sessions of this experiment, and our findings motivated the replication of the analysis using a more traditional multi-trial free recall experiment (Experiment 2). Both studies (Experiments 1 and 2) were approved by the Institutional Review Board at the University of Pennsylvania, and all subjects provided informed consent to participate in the studies.

Experiment 1.

As Experiment 1 appeared in a prior report, we only provide a brief overview of those methods, referring the reader to the original paper for a complete description (Katerman et al., 2022). Katerman et al. presented subjects (N = 57) with a series of 576 common nouns. As soon as each word disappeared, subjects had to say the word aloud following a brief (1 sec) delay (see Figure 1A). Subjects performed five sessions of this task, with the same set of 576 words appearing in a new random order each session. These 576 words corresponded to the same 576-word pool as used in the PEERS4 study (Aka, Phan, & Kahana, 2021; Kahana, Aggarwal, & Phan, 2018; Weidemann & Kahana, 2021). At the beginning of the sixth session, the experimenter surprised subjects by asking them to freely recall as many of the 576 items as they could remember during a 10-minute retrieval period, while also vocalizing any words that came to mind, a version of the task known as externalized free recall (EFR) (Kahana, Dolan, Sauder, & Wingfield, 2005; Lohnas, Polyn, & Kahana, 2015; Zaromb et al., 2006). [1] Following this initial recall task, subjects again saw each of the 576 items (in a new random order) and attempted to immediately recall each item after a brief 1 sec pause. At the start of each of the next four sessions, subjects

[1] Unlike earlier variants of externalized free recall, we did not ask subjects to press a key following responses to indicate whether they thought it was not one of the items on the target list.

performed the same initial free recall task, attempting to recall words studied (and, in some cases, recalled) on previous sessions. Following this initial recall task, they performed the same immediate recall task involving each of the 576 items. Because the initial free recall task occurred at the beginning of sessions 6 through 10, subjects had five trials in which they attempted to recall a very large set of words following a delay of at least one day and often several days. By having subjects recall a large number of common English words after a very long delay this task forced them to distinguish between words experienced in a specific context (of the lab) and those experienced during their daily activities. As in (Katerman et al., 2022), we analyzed data from subjects ($n = 40$) who completed seven or more sessions and met a minimum performance criterion (see Katerman et al., for details).

Experiment 2.

Seven young adults (ages 18-35, four males, three females, six reported right-hand dominant), recruited among the students and staff at the University of Pennsylvania, each contributed 10 sessions of a free recall experiment featuring a final free recall task. Two sessions from two different subjects were not included in the analysis due to technical failure to save the collected data.

In each of the 10 sessions subjects studied and freely recalled 26 word lists, each comprising 12 unique items, drawn from a pool of 312 common nouns used in several prior studies (Ezzyat et al., 2018)[2]. In each list, six words appeared three times, three words appeared twice and three words appeared just once, totalling 27 word presentations per list (see Figure 1B). An algorithm determined placement of repeated and once presented items into the 27 positions of each list with the goal of keeping repeats well separated for the benefits of the spacing effect, while balancing this with a very low predictability for word order or whether a particular word will be a repeated one. The algorithm also ensures that single presented items appear in an unbiased ordering so that the word presentation

[2] `https://openneuro.org/datasets/ds004789/versions/3.1.0`

position for non-repeats is uniformly distributed throughout the list. The list generation algorithm generates and scores 500 permutations of word orderings of double and triple presentations, retrying each one up to 20 times if any repeated words are placed adjacent to each other. Then, all 500 remaining word orderings are given a proximity score p:

$$p = \sum_{i=1} \frac{1}{d_i - 1} \tag{1}$$

where d is the list spacing between all repeat pairings i (where for thrice repeated words, this means between the first and second, and between the second and third presentations), and where any surviving lists with adjacent repeats have a proximity score of infinity. Then the lowest proximity scoring word ordering is selected, corresponding to the list with the most spaced out repetitions out of the 500 generated lists. Single presentation words are then distributed throughout the list in a uniform unbiased manner, and the resulting ordering is used for presentation.

All sessions started with an instructions video followed by a practice list. We included the practice trial in our analyses, which resulted in a total of 312 unique words distributed across 26 lists. Before the start of each list (including the practice list), a fixation cross appeared on the screen for 7s, followed by a 3-s visual countdown video alerting subjects of imminent list onset. After a jittered interval (0.75-1s), each list item appeared as white text on a black background for 1,600 ms and was followed by a jittered interstimulus interval of 750–1,000 ms (uniformly distributed). Following the final interstimulus interval, a fixation cross appeared on the screen for 7s, and then a tone sounded and a row of asterisks appeared onscreen for 1s, indicating the start of a 45s free recall period. The end of the 45-second free recall period was also marked by an audible tone. In addition to recalling as many words from the just studied list, the instruction video also asked subjects to say out loud any word that came to mind whether or not it was on the most recent list, a version of the task known as externalized free recall (Kahana et al., 2005; Lohnas et al., 2015; Zaromb et al., 2006)[3]. Following recall of the final list,

[3] Unlike earlier variants of externalized free recall, we did not ask subjects to press a key following

text on the screen instructed subjects that they had 10 minutes to freely recall as many words as they can from the lists they studied during the session (final free recall; e.g., Craik (1970) and Kuhn et al. (2018)), as well as any other words that come to mind. After subjects read the instructions, a row of asterisks, accompanied with a tone, appeared on the screen for 1s, indicating the start of the 10-minute final free recall period.

Subjects performed 10 sessions of this task completed on separate days. Because each session involved studying and attempting to recall the same set of 312 words (randomized separately in each session), memory of the items experienced in earlier sessions could potentially serve as a source of facilitation and/or interference in subsequent sessions. As in previous studies carried out in the senior author's lab, a computer recorded subjects spoken recalls for offline annotation (Solway, Geller, Sederberg, & Kahana, 2010).

Data and Code Availability

Data for Experiment 1 may be downloaded from `https://openneuro.org/datasets/ds004395/versions/2.0.0`. Data from Experiment 2 and code for both experiments may be downloaded from: `https://github.com/daniellarafla/RaflEtal24/`. This study was not preregistered.

General Analysis Methods

Despite important structural differences between these two experiments, they both require subjects to repeatedly study a long list of words over 10 sessions, and feature a 10-minute retrieval period where subjects attempt to recall over a hundred words studied and recalled on previous sessions. Our analyses focus on the 10-minute retrieval period (cumulative free recall) of both experiments with the goal of understanding how memory organization for a long list of words changes with repeated learning trials across multiple days. We will examine the main factors that shape the organization of memory during free

---

responses to indicate whether they thought it was not one of the items on the target list

recall - semantic, temporal and subjective organization (Howard & Kahana, 2002; Kahana, 1996; Tulving, 1962). We examine the overall effects as well as their evolution with learning across sessions.

Semantic clustering

We follow standard practice by representing words as high-dimensional vectors and estimate the semantic similarity of two given words as the cosine similarity between their associated vectors. We derive these vectors from a Word2Vec 300-dimensional continuous Skip-gram model trained with negative sampling on the Google News dataset (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, G., & Dean, 2013) [4].

Following the method described by Howard and Kahana (2002), we binned all word pairs of each word pool into deciles based on their similarity scores, then computed the probability of making recall transitions to these bins.

To understand how semantic organization evolves with learning across sessions, we needed a way to quantify the magnitude of the semantic clustering effect for each session. We do this by computing a semantic clustering score for each session following the methods described in Polyn, Norman, and Kahana (2009). For any given session, we computed a percentile score for each recall transition made during cumulative free recall. This percentile score ranked the similarity between any pair of successively recalled words relative to the similarity between the first of the two words and all the other words that the subject could have recalled next. We then average the percentile scores across all recall transitions made during the session in question to obtain a semantic semantic clustering score characterizing it. Finally, we average those scores across subjects to obtain a single semantic clustering score for each session.

---

[4] downloaded from: https://code.google.com/archive/p/word2vec/

Temporal clustering

We examine how the temporal proximity of words at encoding affects the organization of recall by computing the lag-CRP function developed by Kahana (1996). The function computes the probability that two items separated by a given temporal distance (lag) during list presentation will be successively recalled during the retrieval period. The temporal distance, or lag, between two words corresponds to the number of items separating these words at study. Kahana (1996) computed these conditional probabilities by dividing the number of observed transitions of a given lag by the number of possible transitions of that lag.

Typically, researchers have computed the lag-CRP curve for free recall events following one pure list. This is not the case for the cumulative free recall events analyzed in this study. In Experiment 2 subjects attempt to recall items studied over multiple lists. We deal with the list partitioning of studied items by only taking into account transitions made between items that were studied in the same word list. That is, we compute the CRP for any given lag based on each list separately, and then average these probabilities across lists. Moreover, the repetition of list items up to three times in word lists of Experiment 2 further complicates the analysis by introducing the possibility of multiple lags between any two words. We adapt the lag-CRP function to this situation following the method used by Kahana and Howard (2005) whereby the lag between any two words with multiple serial positions corresponds to the minimum lag between those words. It is also important to note that in Experiment 2, the most recent item presentation phase based on which we compute the lag between any two words immediately precedes cumulative free recall. In Experiment 1, the most recent item presentation phase occurs one session before the cumulative free recall period analyzed.

Finally, to quantify the magnitude of temporal clustering for any given session, we compute the temporal clustering score as described in Polyn et al. (2009). For each recall transition we compute the percentile rank of the temporal distance at encoding (lag)

between the successively recalled words relative to all the other words that the subject could have recalled next. We average the percentile ranks aross all recall transitions to obtain a temporal clustering score for each session.

Subjective clustering

The methods we describe in this section for the quantification of subjective organization differ from the standard methods used in the literature. The main methods used to measure subjective organization are the Subjective Organization (SO) method proposed by Tulving (1962) and the Intertrial Repetition (ITR) measure proposed by A. K. Bousfield and Bousfield (1966); see Sternberg and Tulving (1977) for a review of both methods. SO and ITR (as well as the methods derived from them) base their measure of subjective organization on a tally of repeated word sequences (typically adjacent word pairs) across trials. The methods we use shift the focus from word sequences and their persistence across trials to the distances between recalled words and how those distances (or subjective lags) on a given trial modulate recall transitions on the subsequent trial.

First, we use the lag-CRP function to analyze the subjective organization of memory during cumulative free recall. Here, subjective organization is quantified by the degree to which subjects cluster their recalls (both correct and false) as a function of the temporal distance between those recalls during the most recent retrieval period. Specifically, the function computes the conditional probability that two items separated by a given lag during the previous cumulative free recall period will be recalled in succession. The lag between any two recalls corresponds to the absolute distance between them during the previous cumulative free recall period, and in the case where a word is repeated during recall, the lag between this word and any other recalled word corresponds to the minimum lag between them (Kahana & Howard, 2005). In both experiments, the most recent cumulative free recall period based on which we compute the subjective lag between any two items occurs one session before the cumulative free recall period analyzed. We

restricted the analysis to transitions with absolute lags ranging from 0 to 10 and excluded sessions whose prior recall period contained 10 or fewer recalls. As a result, three sessions were eliminated from the data in Experiment 1.

Then, we quantify the magnitude of subjective clustering for each session following Lohnas's (in press) recent adaptation of the standard temporal clustering score (Polyn et al., 2009) to subjective clustering. Lohnas calculates a temporal clustering score with the difference being that the lag between any two recalls corresponds to the absolute distance between them during the most recent retrieval event (as opposed to the distance between them during list presentation). We use this method to obtain a percentile rank of the observed transition averaged across all recalls in a session for each subject to trace the change in subjective clustering over learning trials.

We see that while standard methods limit subjective organization to the persistence of rigid word pair sequences across recall trials, the methods we use in this study are sensitive to the degree to which successive recalls on a given trial were recalled in close proximity during the previous trial. Additionally, both analyses have the advantage of being analogous to the standard methods used to quantify and describe the temporal clustering effect (lag-CRP and temporal clustering score), which allows for a direct comparison of the effects of temporal distance at encoding and retrieval on output order.

Statistical Modeling

For all statistical testing, because some subjects were occasionally missing sessions, we use linear mixed effect models that allow for unbalanced designs. In order to conduct analyses of variance, we fit a model with categorical predictors and varying subject intercepts, which is equivalent to a repeated-measures ANOVA in a balanced design. In order to test for linear relationships and adjust for other variables, we always include all varying intercepts and slopes by subject, i.e. the maximal model. Because this random effects structure might exceed the amount of structure that can be estimated from the

data, we would then iteratively reduce the model by removing singular random effect terms using the procedure described in Bates, Kliegl, Vasishth, and Baayen (2018). For inference in all mixed effect models, we use Satterthwaite (1946) approximations to the F and t distributions (Giesbrecht & Burns, 1985; Fai & Cornelius, 1996).

Results

Our analyses focus on the 10-minute cumulative free recall periods at the start of sessions 6-10 in Experiment 1 and at the end of each session in Experiment 2. Analyzing these periods shows how memory changes with cumulative experience across multiple days. For each experiment, we first asked how the number of correct recalls and intrusions changed as subjects gained experience with the memoranda. We next examined the effect of long-standing semantic associations and temporal associations formed during encoding on the organization of memory during the 10-minute retrieval periods. Finally, we turned to the influence of previous retrieval on the organization of memory by examining the extent to which subjects increasingly stereotype their recalls over trials, including the manner in which they repeat extra-list intrusions across sessions.

Changes in recall performance across sessions

Figure 2A,C show the average number of correct recalls as a function of session number. In both experiments, correct recall rates increased reliably across sessions. A linear mixed effects model predicting the number of correct recalls as a function of session number tested this observation. In both experiments, we found a positive linear relationship between session number and the number of correct recalls (Experiment 1: $\beta = 20.49, t(38.9) = 7.31, p < 0.0001$; Experiment 2: $\beta = 3.82, t(5.16) = 4.38, p < 0.01$).

When recalling a large set of items after a delay one might expect a higher rate of recall errors (intrusions) than in standard free recall (e.g., Katerman et al., 2022). Here, however, Experiments 1 and 2 differed dramatically in the delay interval: In Experiment 1, subjects attempted cumulative free recall after one or more days, during which they

experienced many other linguistic stimuli, whereas, in Experiment 2, they did so at the end of each session. Figure 2B,D show the number of extralist intrusions across sessions in each experiment. In Experiment 1, subjects made a large number of intrusions which significantly increased from Sessions 6 to 10 (as revealed by a model predicting intrusions as a linear function of session number, $\beta = 3.62$, $t(40.40) = 3.69$, $p < 0.001$). However, subjects made many fewer intrusions in Experiment 2, and their intrusion rates did not consistently change across sessions ($\beta = -0.069, t(6.49) = -1.36, n.s.$).

Changes in semantic and temporal organization across sessions

Examining the change in recall performance across learning trials reveals patterns of knowledge accumulation, but it does not reveal the changes in the organization of memory retrieval underlying those patterns. To understand the organization of memory, researchers have typically studied the phenomenon of clustering, defined as the tendency to group words that are associated along a given dimension during recall. Free recall analyses have typically focused on clustering based on the semantic similarities among list items (semantic clustering) and on the temporal order in which those items were presented (temporal clustering). In this section, we assess the effects of semantic and temporal proximity of items on output order, and examine the change in these effects across sessions in each experiment.

Figures 3A,D illustrate semantic organization during cumulative free recall in Experiments 1 and 2 respectively. Each figure shows the conditional response probability (CRP) as a function of semantic similarity, as measured by cosine similarity of trained Word2Vec vectors (Mikolov, Chen, et al., 2013), during the cumulative free recall period. The values on the x-axis represent the means of similarity scores within ten discrete bins representing deciles of the distribution of similarity scores. Figures 3B,E illustrate the contiguity effect as seen in the conditional response probability as a function of lag during encoding, or lag-CRP curve (Kahana, 1996). Subjects in Experiment 1 exhibited no

contiguity effect based on the order in which they most recently encoded items (Figure 3E). However, subjects in Experiment 1 appear to have relied on the semantic associations between items to generate recalls, as evidenced in the strong semantic proximity effect in panel D. In contrast, Experiment 2 exhibited both strong semantic and temporal clustering (panels A, B) during cumulative free recall, consistent with prior studies of final free recall (Kuhn et al., 2018).

We then turn to the change in these effects across sessions. Figure 3C,F show the change in the semantic and temporal clustering scores (Polyn et al., 2009) across sessions in Experiments 1 and 2, respectively. To examine whether there was a change in these temporal scores across sessions, we conducted four analyses of variance, one for each effect and experiment. In Experiment 1, temporal clustering score does not significantly vary as a function of session ($F(4, 144.94) = 0.64, n.s.$), but the ANOVA revealed a significant main effect of session on semantic clustering score ($F(4, 145.16) = 6.44, p < 0.0001$). We followed up on this result with a model predicting semantic clustering score as a linear function of session number. The model indicates that semantic clustering score increases across sessions in Experiment 1 ($\beta = 0.022, t(40.13) = 3.92, p < 0.001$). In Experiment 2, there was no significant main effect of session on semantic ($F(9, 52.01) = 0.96, n.s.$) or temporal clustering ($F(9, 52.14) = 0.20, n.s.$).

Since subjects also increase the number of correct recalls across sessions in Experiment 1 (Figure 2A), the increase in semantic clustering across sessions could be a consequence of the increased number of possible clustering opportunities and not an independent effect of experience in the task. To separate out the effects of time in the task and recall performance on semantic clustering score, we fit a model with three covariates: session number (as a linear predictor), the subject-level mean number of correct recalls, and the session-level number of correct recalls relative to the subject-level mean. Both session number ($\beta = 0.018$, t(55.34) = 2.18, p < 0.05) and the subject-level mean number of correct recalls ($\beta = 0.047$, t(37.63) = 4.67, p < 0.0001) significantly predicted semantic

clustering score in Experiment 1. The session-level number of correct recalls did not significantly predict semantic clustering score within subject ($\beta = 0.009, t(74.40) = 0.84, n.s.$). These results indicate that the increase in semantic clustering is more likely an independent effect of learning over time in the task rather than a consequence of enhanced performance. Moreover, this analysis shows a positive correlation between the overall mnemonic abilities of the subject and the subject's tendency to semantically organize their recalls.

Changes in subjective organization across sessions

While temporal and semantic clustering make up the bulk of the literature about the organization of memory during free recall, a number of studies examined the subjective organization of memory when subjects learn the same word-list in different orders over multiple learning trials. These studies typically use lists of unrelated words to quantify the degree to which subjects impose their own organization to the input material, and report increasingly stereotyped responses across learning trials. While reports of subjective organization typically examine this phenomenon across learning trials that occur within a single session, the current study examines subjective organization and its change across learning trials (here, sessions) separated by multiple days.

Figure 4A,C show, for Experiments 1 and 2 respectively, the conditional response probability (CRP) of making recall transitions among words as a function of their distance (or subjective lag) during the free recall period of the previous session. Interestingly, the subjective lag-CRP curves of both experiments exhibit a strong contiguity effect, meaning that the probability of recalling two items in successive output positions increases as a function of how close the output positions of those items were during the previous recall period.

Figure 4B,D show the change in the subjective clustering score across sessions. We measure subjective clustering score by applying the same percentile rank method used to

measure temporal clustering score described in Polyn et al. (2009), and adapted to the measurement of subjective clustering by Lohnas (in press). This method measures the degree to which recalled items are sequentially closer in distance to each other based on the order in which they were recalled a session prior (as opposed to the order in which they were originally presented). Graphs for both experiments qualitatively show a moderate increase in subjective clustering across sessions. To examine whether there was a significant increase in subjective clustering across sessions, we fit mixed effect models predicting subjective clustering as a linear function of session number across subjects to the data from each experiment. The models revealed a significant correlation between subjective clustering and session number in Experiment 1 ($\beta = 0.031, t(38.66) = 3.01, p < 0.01$), and Experiment 2 ($\beta = 0.012, t(6.46) = 4.30, p < 0.01$), suggesting that the difference across sessions is at least partly attributable to learning across sessions.

Because the number of recalls also increases across sessions (Experiment 1: $\beta = 24.05, t(39.15) = 7.36, p < 0.0001$; Experiment 2: $\beta = 3.75, t(5.21) = 4.3, p < 0.01$), one possibility is that this increase is simply due to the increase in the total number of recalls. We fit mixed effect models to the data of both experiments predicting the session-level subjective clustering as a linear function of session number, subject-level number of recalls averaged across sessions, and the session-level number of recalls relative to the subject-level mean. The session-level number of recalls was not a significant predictor of subjective clustering in either experiment (Experiment 1: $\beta = 0.014, t(118.22) = 0.834, n.s.$; Experiment 2: $\beta = -0.006, t(50.07) = -0.65, n.s.$), but subjective clustering score significantly increased as a function of session number in both experiments (Experiment 1: $\beta = 0.027, t(118.72) = 2.22, p < 0.05$; Experiment 2: $\beta = 0.012, t(50.06) = 4.56, p < 0.0001$). Experiment 1 additionally showed a significant effect of subject-level mean number of recalls in predicting the session-level subjective clustering score ($\beta = 0.041$, t(47.6) = 3.73, p < 0.001). This effect was not significant in Experiment 2 ($\beta = 0.007, t(5.02) = 0.25, n.s.$).

Because subjective organization is concerned with the preservation of the order in

which subjects preferentially recall words across sessions, the analysis reflects associations that could be temporal (based on the temporal order in which items were studied), or semantic. Accordingly, we asked whether or to what extent the robust effect of subjective clustering observed in both experiments is associated with the effects of semantic and temporal clustering. For both experiments, we started by estimating a mixed effects model predicting the session-level subjective clustering score as a function of the subject-level mean semantic clustering score, and the session-level semantic clustering score relative to the subject-level mean (M1).

Experiment 1 showed a significant effect of both session-level ($\beta = 0.035, t(84.64) = 3.04, p < 0.01$) and subject-level semantic clustering ($\beta = 0.025, t(45.69) = 2.13, p < 0.05$) on subjective clustering, suggesting that the increase in subjective clustering at least partially reflects the increase in semantic clustering across sessions. However, it may be that both types of clustering increase with more experience with the wordpool rather than one causing the other. We fit a second model that included the number of completed sessions as an additional predictor of subjective clustering (M2). The model including both variables suggests that the amount of experience significantly predicts subjective clustering ($\beta = 0.025, t(133.1) = 2.71, p = 0.01$), while session-level semantic clustering is marginally significant ($\beta = 0.022, t(62.96) = 1.78, p = 0.08$). Subject-level semantic clustering remains a significant predictor of subjective clustering ($\beta = 0.034, t(43.80) = 2.78, p < 0.01$). Therefore, the relationship between semantic and subjective clustering may be simply due to the fact that both increase with experience, rather than the increase in subjective clustering being caused by the increased semantic clustering.

Since semantic clustering remains constant while subjective clustering increases across sessions in Experiment 2, we did not expect session-level semantic clustering to be a significant predictor of subjective clustering in either M1 or M2. This was in fact true of the estimates (M1: $\beta = 0.009, t(15.9) = 0.811, n.s.$; M2:

$\beta = -0.002, t(16.44) = -0.21, n.s.$). However, both models reveal subject-level semantic clustering (M1: $\beta = 0.063, t(4.61) = 4.83, p < 0.01$; M2: $\beta = 0.061, t(4.37) = 5.09, p < 0.01$) to be a significant predictor of subjective clustering. M2 additionally showed that subjective clustering significantly increases as a function of session number ($\beta = 0.012, t(52.16) = 4.58, p < 0.0001$).

Similarly, we estimated two more mixed-effect models (one for each experiment) predicting subjective clustering as a function of the subject-level mean temporal clustering score, the session-level temporal clustering score relative to the subject level, as well as the number of completed sessions. Experience with the task was a marginally significant predictor of subjective clustering in Experiment 1 ($\beta = 0.018, t(121.5) = 1.910, p = 0.058$), but a significant predictor in Experiment 2 ($\beta = 0.012, t(50.79) = 4.54, p < 0.0001$). Given the absence of an overall temporal clustering effect in Experiment 1, it is reasonable to think that the subjective organization of memory could not reflect the temporal organization of words at encoding. Indeed, the model shows that the subject-level mean temporal clustering score does not significantly predict subjective clustering ($\beta = 0.008, t(36.12) = 0.835, n.s.$). Surprisingly however, the session-level temporal clustering score significantly predicted subjective clustering ($\beta = 0.043, t(50.53) = 3.99, p < 0.001$). One potential explanation for this is that both temporal and subjective clustering in Experiment 1 refer to events from the previous session. To the extent that subjects are able to recollect that session, subjects would show both temporal and subjective clustering. In Experiment 2, neither the subject-level mean temporal clustering score ($\beta = -0.004, t(4.97) = -0.13, n.s.$) nor the session-level temporal clustering score ($\beta = -0.007, t(53.99) = -0.50, n.s.$) significantly predicted subjective clustering.

False recall

Our analysis of subjective clustering demonstrated that across multiple days, the output order of subjects' cumulative recall became increasingly stereotyped. We performed these analyses on all recalled items irrespective of whether the items were actually in the set of studied words. Whereas subjects in Experiment 2 rarely committed false recalls, in Experiment 1 subjects made a large number of extralist intrusions and they repeated many of those intrusions across sessions. This suggests that intrusions become an integral part of the increasingly stereotyped memory subjects form across sessions. We now ask whether the repetition of intrusions during retrieval follows a particular trend over time. More specifically, we sought to determine whether the repetition of intrusions followed a pattern of recency (as opposed to being randomly distributed) across sessions.

Figure 5 shows the probability of repeating an intrusion (for the first time) as a function of the number of sessions elapsed since its first occurrence (or session lag). We computed this probability by dividing the number of times an intrusion was first repeated after a given session lag by the number of times this intrusion could have been first repeated after this session lag, conditional on the availability of this lag. Figure 5 shows that the probability of repeating an intrusion at a given session decreased with the number of sessions elapsed between the session where a subject first made the intrusion and the session where they first repeated it.

We conducted an analysis of variance to test for a main effect of session-lag on number of intrusions, which was significant ($F(3, 114.22) = 53.94$, $p < 0.0001$). We then conducted post hoc pair-wise t-tests across all session lags to determine whether the probability of repeating intrusions significantly differed among lags. We find that all comparisons with the first lag were significant, meaning that the probability of repeating an intrusion was significantly greater one session after it was first made compared with two ($M = 0.16$, $t(111) = 9.02$, $p < 0.0001$), three ($M = 0.19$, $t(113) = 10.9$, $p < 0.0001$), and four ($M = 0.19$, $t(114) = 10.6$, $p < 0.0001$) sessions later.

In contrast, subjects made a small number of intrusions in the final free recall task of Experiment 2 (Figure 2D), which prevented replication of the intrusion analysis to the experiment. The fewer intrusions made in Experiment 2 could be explained by the short study-test delay, which made it easier for subjects to retain the words they had just committed to memory with minimal interference from pre-experimental words.

Discussion

In typical studies of episodic memory, subjects attempt to recall memoranda experienced in a target list within a single session. Here we examined two experiments in which recall was assessed repeatedly across multiple days. In both experiments, subjects attempted to freely recall a large corpus of words (576 in Experiment 1 and 312 in Experiment 2). They experienced these words either during encoding sessions one or more days before the cumulative free recall test (Experiment 1) or during a series of study-test trials during the same experimental session (Experiment 2). Each experiment provided a unique window into the evolution of episodic memory across multiple days across which subjects had repeated opportunities to encode all of the words in the corpus. Although experimental psychologists have long investigated the acquisition of knowledge across periods of extended practice, the present experiments offered a unique view of how humans acquire such knowledge under highly unstructured conditions; because we randomized lists on every trial and session, no two encoding trials sought to reinforce the same associations.

Our two experiments differed critically in the instructions during encoding and the delay between ending and recall. In Experiment 1, subjects were only instructed to repeat each word after a brief 1-2 second delay, and their cumulative recall was evaluated after one or more intervening days in which they likely experienced many of the same words outside of the laboratory. In Experiment 2, subjects intentionally studied each word for an immediate free recall test. They then attempted their cumulative recall at the end of the same study session. In both experiments, however, subjects attempted cumulative recall of

a very large set of words that all repeated across these encoding tasks on multiple days.

Data from both studies demonstrate learning across days: From the first cumulative recall test to the last one, subjects increasingly recalled greater numbers of experienced items. The long-delay cumulative recall test of Experiment 1 (which entailed incidental encoding) demonstrated much larger gains across sessions than the within-session cumulative recall test of Experiment 2. This difference likely reflected multiple variables that differed across the two studies, including the fact that Experiment 2 yielded far higher levels of recall on Session 1 and would thus have produced a build-up of interference across the recall phase by challenging subjects to remember which words they had and had not already recalled.

In both experiments, we observe strong evidence of semantic organization in cumulative free recall. Still, only Experiment 2, which entailed an immediate free recall phase before cumulative recall, demonstrated temporal organization during cumulative recall. We interpret this difference as arising from the joint effects of variable delays across the experiments and output encoding during recall. In Experiment 1, subjects only had an opportunity to recall the studied words after one or more days, during which they likely experienced massive interference with the 576-word corpus. In contrast, Experiment 2 tasked subjects with immediate free recall which likely further strengthened associations among items studied in similar temporal contexts during learning. Then, cumulative recall was tested at the end of the session, before subjects would have encountered extra-experimental materials and while subjects were in the same spatiotemporal context of the experimental session. Across sessions of Experiment 1, subjects exhibited increasing levels of semantic organization of the memoranda, which aligned with their dramatic increase in recall performance across sessions. We next consider the central question of our paper, which concerned the manner in which cumulative recall itself shaped the organization of memory on subsequent sessions. Before discussing those results, however, we wish to remind the reader of classic prior work on subjective organization in multi-trial

free recall.

Subjective Organization and Output Encoding

Over a half-century ago, Tulving captured the imagination of experimental psychologists by suggesting that a subject's ability to form unique organizational structures of disorganized memoranda underlies successful learning (Tulving, 1962, 1968; Sternberg & Tulving, 1977). Tulving's ingenious demonstration utilized the well-established method of multi-trial free recall. Subjects would attempt to learn a list of randomly selected words by repeatedly studying them and freely recalling them, with the order of study randomized from trial to trial. Across trials, the order of recalled words became increasingly stereotyped, and the degree of this stereotypy – which Tulving labeled subjective organization - tracked the rate at which subjects learned the list. As learning a single list transpires within one experimental session, classic studies of subjective organization tracked the evolution of memory organization within a brief span of time (Kahana & Wingfield, 2000). But just because organization and learning occur in tandem, one cannot implicate one as causing the other. Indeed, numerous researchers have noted that both organization and learning may result from a common cause (Crowder, 1976; Kahana, 2012).

In both multi-trial free recall and multi-session cumulative recall, repeated presentation of randomized lists will obscure the temporal structure defined by the contiguity relations among items. As this occurs, the effect of the preexperimental semantic relations may gain prominence and even dominate over temporal organization as defined by the most recent presentation order. However, output encoding provides another channel that would lead to subjective organization, as seen in the increasing resemblance between output order on successive cumulative recall trials. Specifically, if successively recalling a pair of items on trial i leads to associative learning (e.g., by associating the recalled items with similar contextual representations or forming a direct item-to-item association) then subjects will be more likely to emit the same sequence on the next trial.

To measure subject organization, we adapted the now well-established lag-CRP measure of temporal organization. Specifically, we related the probability of successively recalling item i and item j on a given trial to the temporal lag between i and j in the previous trial's output sequence. This metric revealed strong evidence for subjective organization, as seen in the increasing tendency to exhibit consistent trial-to-trial clustering, in both Experiments (see Figure 4). These findings indicate that the order of retrieval plays a crucial role in the subsequent organization of memory, especially when recall consistently occurs after a long delay.

Analyses of recall errors (intrusions) provide further evidence for the critical role of prior retrieval in shaping subsequent recollection. Experiment 1 instructed subjects to adopt a lax criterion for recall, asking them to say aloud any word that came to mind during the cumulative free recall phase. Because of the very long delay between prior encoding and free recall (usually several days), subjects made many intrusion errors. This, in turn, enabled us to unveil important trends pertaining to the accumulation of false recall with learning over multiple days. We found that subjects repeated a large proportion of their intrusions on at least one subsequent session. Further, we show that when subjects make an intrusion, they are more likely to repeat it for the first time in the session immediately following the session where they first made it (compared with two, three, or four sessions later). This finding substantially extends prior demonstrations of recency for prior-list intrusions. Whereas prior studies demonstrate a recency effect for prior-list intrusions across multiple lists (Zaromb et al., 2006) we demonstrate a recency effect for extra-list intrusions across multiple days.

While we adopt output encoding as an explanation for increased subjective clustering over sessions, one alternative explanation is that, as subjects recall more items over sessions, they learn to organize them according to pre-existing semantic associations, resulting in the appearance of similar organization across sessions. In support of this explanation, one could point out that both semantic and subjective clustering increase

across sessions in Experiment 1. In addition, subjects tend to repeat intrusions across sessions and these intrusions tend to be semantically related to words studied during the encoding period (Zaromb et al., 2006). However, when we account for subjects' time in the task and overall tendency to semantically organize their recalls, we find that semantic clustering does not predict subjective clustering at the session level. It is possible, though, that subjects have idiosyncratic personal semantic associations that population level models like Word2Vec do not capture. We argue that two key results are challenging to explain without output encoding. One, we demonstrate that the repetition of these intrusions exhibits a strong recency effect: Intrusions are more likely to be repeated for the first time in the session immediately following the session of their first occurrence relative to two, three, or four sessions later. Semantic associations may be repeated but, without the recalled items being themselves encoded, there is no reason for them to be more likely to be repeated in a closer session. Two, despite the lack of an overall temporal clustering effect in Experiment 1, we find that when we account for time in the task and overall temporal clustering, temporal clustering at the level of the session predicts subjective clustering. Crucially, under the output encoding assumption, both measures rely on temporal associations created during the previous encoding period. This correlation is not present in Experiment 2, where temporal clustering and subjective clustering refer to different sessions. Thus, while this evidence does not completely rule out the possibility that the subjective clustering effect reflects pre-experimental subjective associations, it aligns more closely with the view that subjects encode and learn associations they make during retrieval.

Relation to Hypermnesia and the Testing Effect

Because our experiments included both encoding and retrieval phases, we cannot know how the organization of recall would have evolved differently if subjects were only asked to perform cumulative recall, without the interpolated item encoding phase. Studies

of hypermnesia provide some insights into the changes in organization over repeated testing (Roediger & Challis, 1989). Indeed, a recent study by Lohnas (in press) found that during final free recall, transitions among same-list items that subjects also recalled on an initial test exhibited greater temporal and semantic clustering during final recall. Lohnas also examines the extent to which subjects preserve their subjective recall sequences on successive tests when these tests are not separated by additional encoding. Consistent with our findings, Lohnas found that subjects show a strong tendency to preserve their recall order between tests. Given that subjects were tested twice on the same items without an intervening encoding phase, these findings provide another source of evidence for output encoding during the first recall test.

Studies of the testing effect also offer evidence for learning during test trials. For example, Karpicke and Roediger (2007) show that delaying an initial test enhances long-term retention, regardless of the temporal spacing of subsequent tests. Accordingly, to the extent that the testing effect and the strong retention of associations made during retrieval are related, one could presume that the difficulty of retrieval under the extreme condition of consistently delaying recall by days explains the strength of the subjective contiguity effect in Experiment 1 relative to Experiment 2. From these inferences emerges an associative view of the testing effect. Subjective organization, enhanced by the difficulty of retrieval caused by a long delay, could help to explain the beneficial effects of testing over restudying.

The subjective contiguity effect reflects the retention of associations made during recall and their subsequent retrieval at rates that at least compete with temporal clustering (Experiment 2) and, at best, far exceed the retrieval rate of studied associations (Experiment 1). This observation suggests that the testing effect could be explained by the formation of stronger associative connections between items during recall relative to study. The existence of a relationship between the testing effect and subjective contiguity effect is further supported by the fact that both effects support long-term retention.

Conclusion

We report analyses of two multi-trial learning experiments that allowed us to examine the dynamics of episodic memory across multiple days, recalling the same large corpus of words (576 in Experiment 1 and 312 in Experiment 2) on repeated cumulative free recall trials. Because subjects repeatedly attempted to recall these very long 'lists,' we could study how the dynamics of prior free recall influenced recall dynamics days later. Extending classic demonstrations of subjective organization (Tulving, 1962, 1966), both of our experiments demonstrated a subjective clustering effect that increased across days. The robustness of subjective clustering and its increase across learning trials attests to the crucial role that retrieval plays in shaping long-term episodic memory for words and events. The present results provide a new source of behavioral evidence for the emerging view that thoughts become memories – a concept long recognized by memory scholars but rarely studied in the laboratory.

References

Aka, A., Phan, T. D., & Kahana, M. J. (2021). Predicting recall of words and lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(5), 765–784. doi: http://dx.doi.org/10.1037/xlm0000964

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. *arXiv*(arXiv:1506.04967). doi: 10.48550/arXiv.1506.04967

Bousfield, A. K., & Bousfield, W. A. (1966). Measurement of clustering and of sequential constancies in repeated free recall. *Psychological Reports*, *19*(3, PT. 1), 935–942. doi: 10.2466/pr0.1966.19.3.935

Bousfield, W. A., Sedgewick, C. H., & Cohen, B. H. (1954). Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology*, *67*, 111-118.

Craik, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*(2), 143–148. doi: https://doi.org/10.1016/S0022-5371(70)80042-1

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Lawrence Erlbaum and Associates.

Ezzyat, Y., Wanda, P., Levy, D. F., Kadel, A., Aka, A., Pedisich, I., ... Kahana, M. J. (2018). Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. *Nature Communications*, *9*(1), 365. doi: 10.1038/s41467-017-02753-0

Fai, A. H.-T., & Cornelius, P. L. (1996). Approximate f-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation*, *54*(4), 363–378. doi: 10.1080/00949659608811740

Giesbrecht, F. G., & Burns, J. C. (1985, June). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results.

Biometrics, *41*(2), 477. Retrieved from http://dx.doi.org/10.2307/2530872 doi: 10.2307/2530872

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*(2), 575–596. doi: 10.1037/a0033715

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. *Brain Mechanisms and Learning*, 37-46.

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*(1), 85–98. doi: 10.1006/jmla.2001.2798

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109. doi: 10.3758/BF03197276

Kahana, M. J. (2012). *Foundations of human memory*. New York, NY: Oxford University Press.

Kahana, M. J., Aggarwal, E. V., & Phan, T. D. (2018). The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1857–1863. doi: https://doi.org/10.1037/xlm0000553

Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *Journal of Gerontology: Psychological Sciences*, *60*(2), 92–97. doi: 10.1093/geronb/60.2.P92

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review*, *12*(1), 159–164.

Kahana, M. J., & Wingfield, A. (2000). A functional relation between learning and organization in free recall. *PBR*, *7*, 516-521.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151-162.

Katerman, B. S., Li, Y., Pazdera, J. K., Keane, C., & Kahana, M. J. (2022). EEG

biomarkers of free recall. *NeuroImage*, *246*, 118748. doi: 10.1016/j.neuroimage.2021.118748

Kirkpatrick, E. A. (1894, November). An experimental study of memory. *Psychological Review*, *1*(6), 602-609.

Klein, K. A., Addis, K. M., & Kahana, M. J. (2005). A comparative analysis of serial and free recall. *Memory & Cognition*, *33*(5), 833–839.

Kuhn, J. R., Lohnas, L. J., & Kahana, M. J. (2018). A spacing account of negative recency in final free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1180–1185. doi: 10.1037/xlm0000491

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. doi: 10.3758/BF03210951

Lohnas, L. J. (in press). Associations supporting items gained and maintained across recall tests. *Journal of Experimental Psychology: Learning, Memory Cognition*.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337–363. doi: 10.1037/a0039036

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arxiv preprint arXiv:1301.3781*. doi: 10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., G., C., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv*. doi: 10.48550/arXiv.1310.4546

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology*, *10*, 465-501.

Murdock, B. B. (1974). *Human memory: Theory and data*. Potomac, MD: Lawrence Erlbaum and Associates.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and

retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. doi: https://doi.org/10.1037/a0014420

Roediger, H. L., & Challis, B. H. (1989). Hypermnesia: Improvements in recall with repeated testing. in current issues in cognitive processes: The tulane flowerree symposium on cognition. In (p. 175-199). New Jersey: LEA inc.

Satterthwaite, F. E. (1946, December). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110. Retrieved from `http://dx.doi.org/10.2307/3002019` doi: 10.2307/3002019

Siegel, L. L., & Kahana, M. J. (2014). A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(3), 755-764. doi: 10.1037/a0035585

Solway, A., Geller, A. S., Sederberg, P. B., & Kahana, M. J. (2010). Pyparse: A semiautomated system for scoring spoken recall data. *Behavior Research Methods*, *42*(1), 141-147. doi: 10.3758/BRM.42.1.141

Sternberg, R. J., & Tulving, E. (1977). The measurement of subjective organization in free recall. *Psychological Bulletin*, *84*(3), 539-556.

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review*, *69*(4), 344–354. doi: 10.1037/h0043150

Tulving, E. (1968). Theoretical issues in free recall. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory* (p. 2-36). Englewood Cliffs, N. J.: Prentice-Hall.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford.

Tulving, E. (1985). How many memory systems are there? *American psychologist*, *40*(4), 385.

Weidemann, C. T., & Kahana, M. J. (2021). Neural measures of subsequent memory reflect endogenous variability in cognitive function. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(4), 641–651. doi: 10.1037/xlm0000966

Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., & Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 792–804. doi: 10.1037/0278-7393.32.4.792

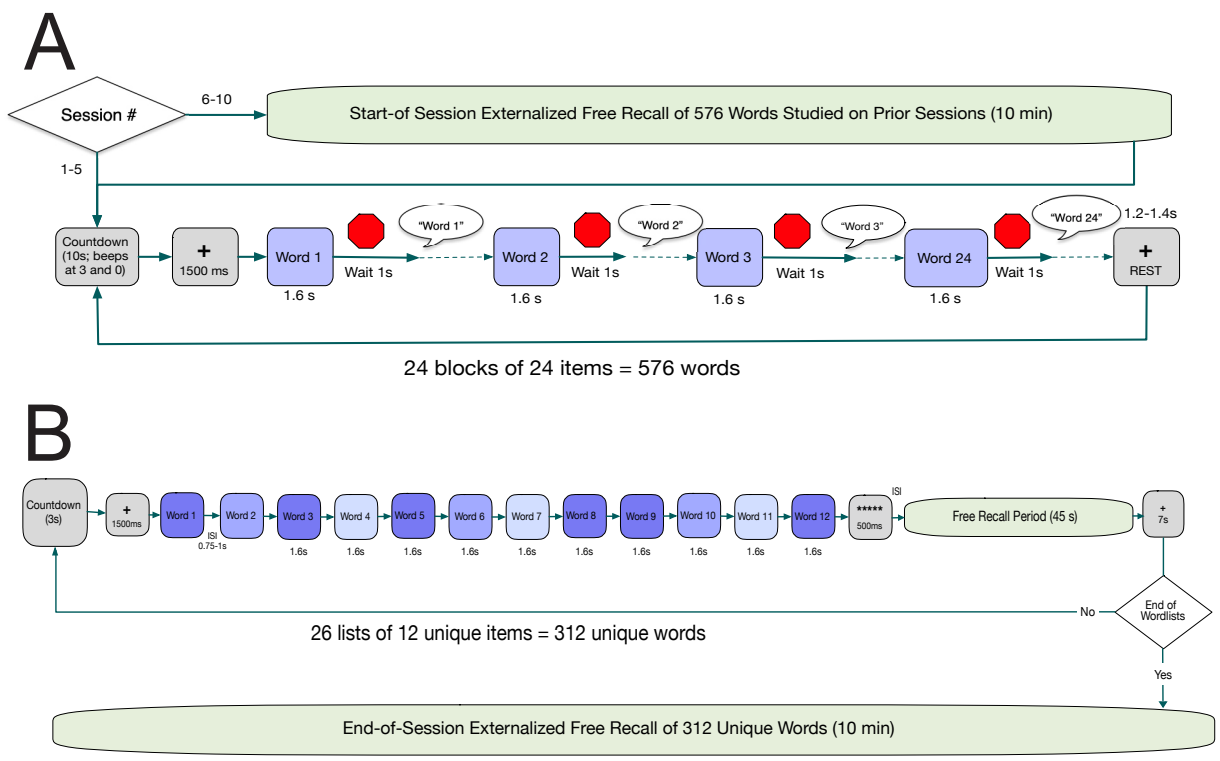

Figure 1

(A) Schematic of Experiment 1. In sessions 1-5 (phase 1), subjects recall each of 576 words one second after it disappears from the screen. The 576 word presentations are divided into 24 blocks of 24 words each. Sessions 6-10 (phase 2) start with a 10-minute externalized free recall period where subjects recall as many of the 576 studied words as they can, while also vocalizing any other words that come to mind. Subjects then proceed with the same immediate recall task of phase 1. (B) Schematic of Experiment 2. Subjects study and then freely recall a series of 26 lists, each consisting of 12 unique word presentations. Different colors for word presentations indicate the number of times each word was presented (one, two or three presentations indicated by light to darker colors respectively). At the end of the session, subjects go through a 10-minute externalized free recall period where they recall as many of the 312 studied words as they can, while also vocalizing any other words that come to mind. Subjects perform 10 sessions of this task on separate days. Each session involved studying and attempting to recall the same set of 312 words which were randomly distributed across lists in each session.

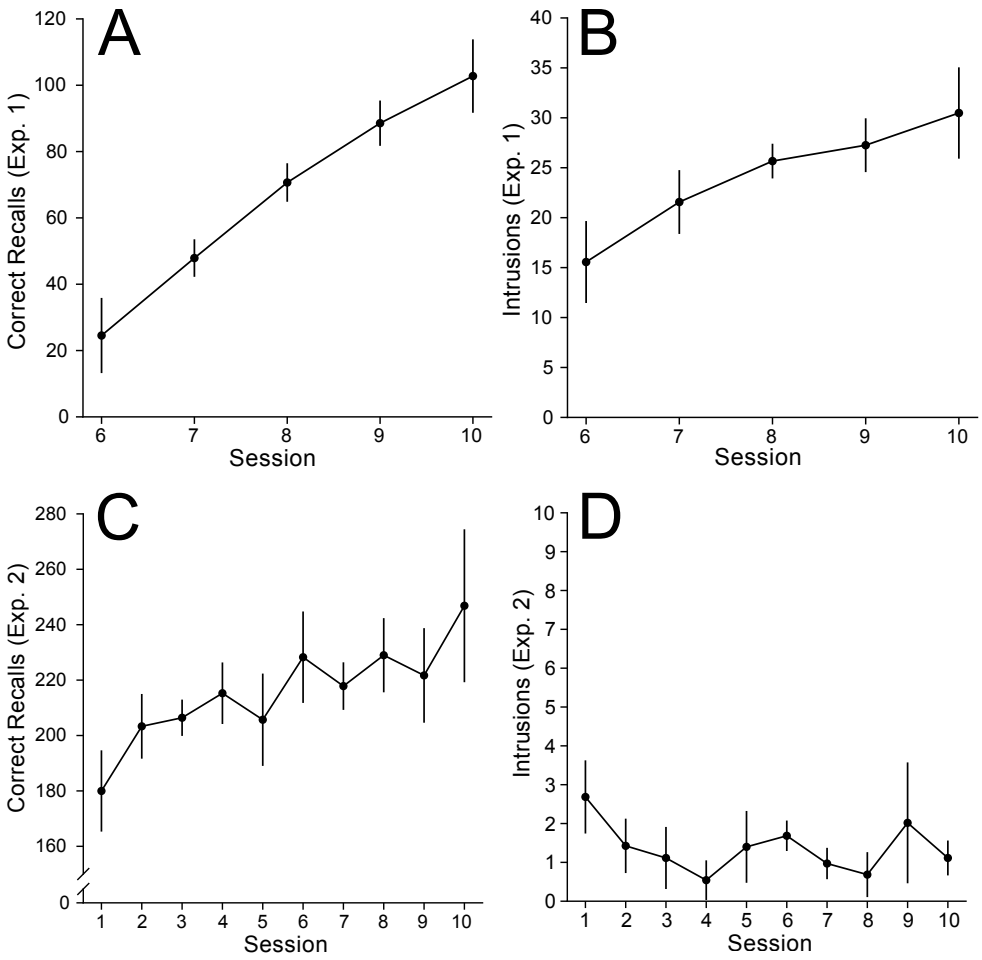

Figure 2

(A) Number of correct recalls as a function of session number in Experiment 1. (B) Number of intrusions as a function of session number in Experiment 1. (C) Number of correct recalls as a function of session number in Experiment 2. (D) Number of intrusions as a function of session number in Experiment 2. Error bars represent 95% confidence intervals. The Loftus and Masson (1994) procedure for computing confidence intervals for within-subject designs was used.

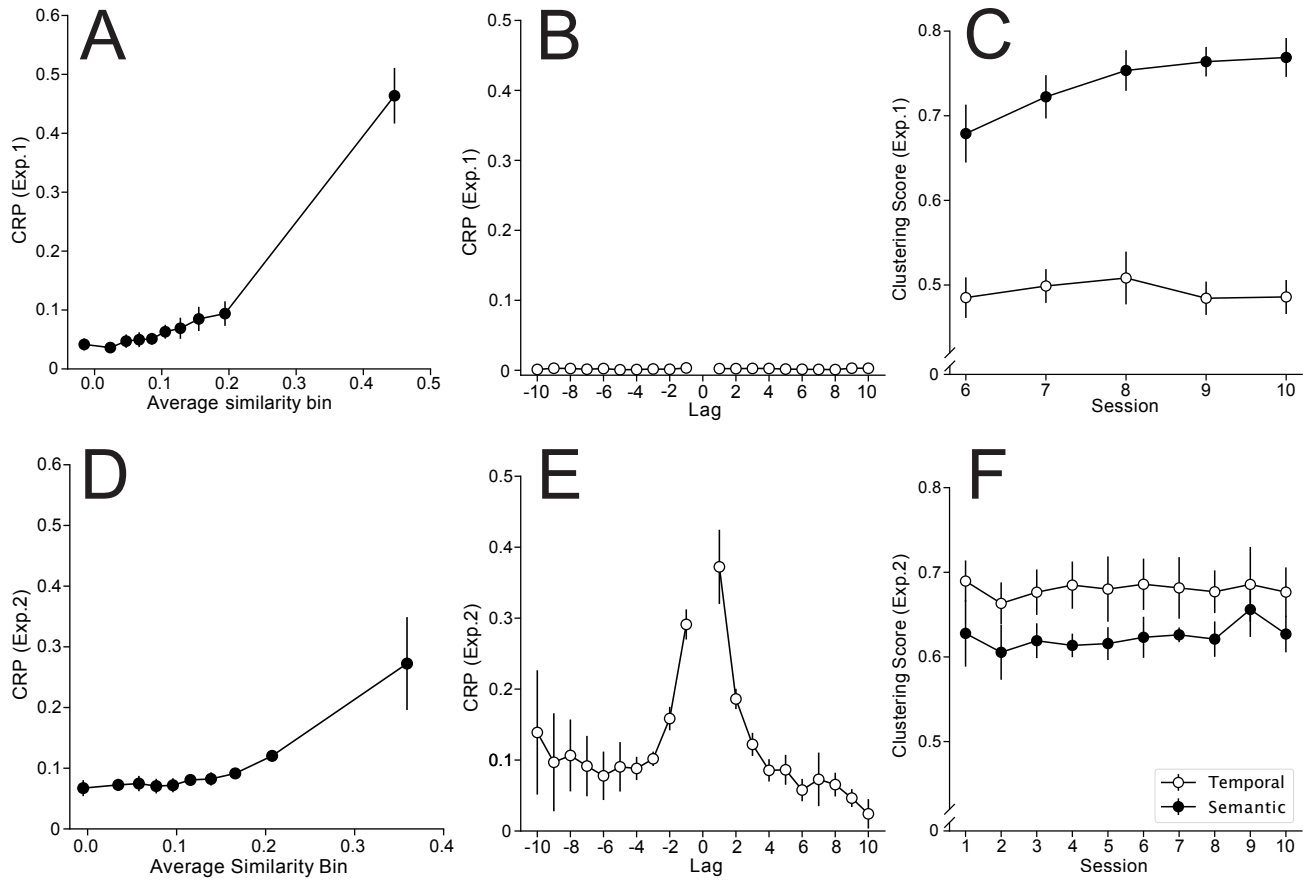

Figure 3

(A) The Conditional Response Probability (CRP) of making recall transitions to words in various semantic similarity score bins in Experiment 1. (B) CRP of successively recalling two items as a function of their temporal distance (lag) during encoding, or lag-CRP curve for Experiment 1. Note that the most recent item presentation phase based on which lags are computed occurs one session prior to the cumulative free recall period analyzed. (C) Semantic and temporal clustering scores as a function of session number in Experiment 1. Clustering scores are calculated following the methods described in Polyn et al., 2009 (see General Analysis Methods). (D) CRP as a function of semantic similarity in Experiment 2. (E) Lag-CRP curve in experiment 2. Note that due to the repetition of items during word presentation in Experiment 1, lag corresponds to the minimum distance between pairs of words at encoding (see Kahana and Howard, 2005). (F) Semantic and temporal clustering scores as a function of session number in Experiment 2. All error bars represent 95% confidence intervals. Error bars in panels C and F used the Loftus-Masson (1994) adjustment for between-subject variability.

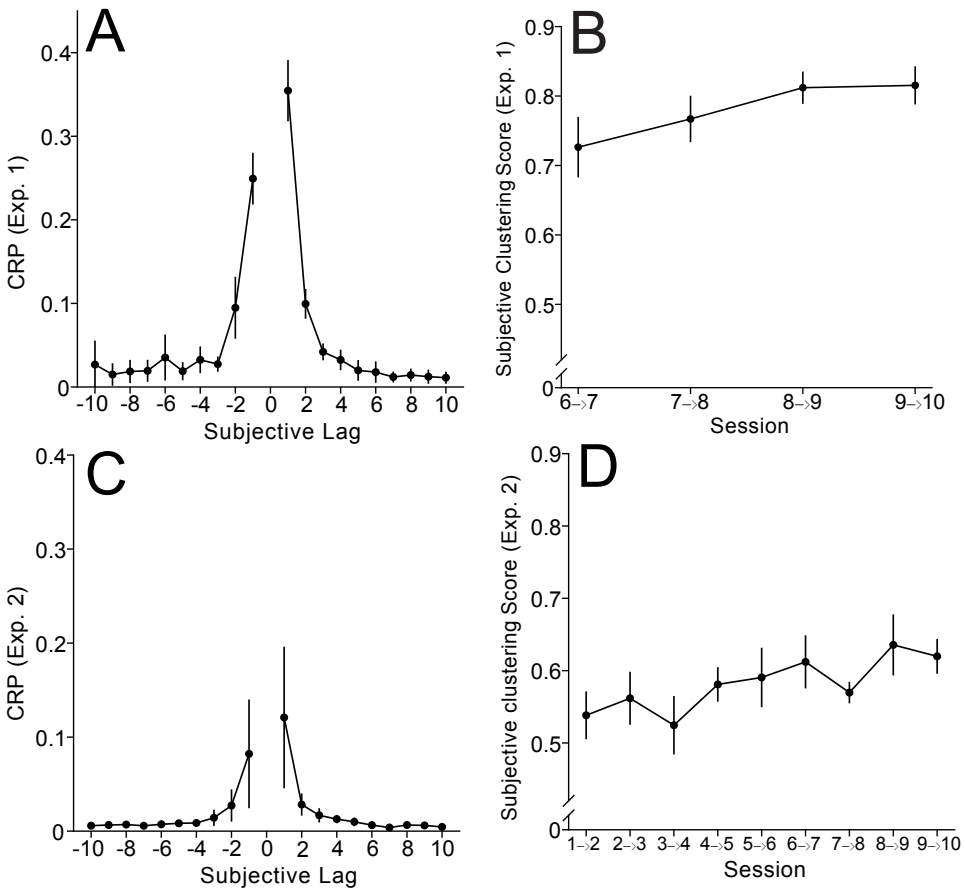

Figure 4

(A) The Conditional Response Probability (CRP) of successively recalling items as a function of their temporal distance during previous retrieval, or subjective lag-CRP for Experiment 1. Note that due to the subjects' tendency to repeat words during free recall, the lag between two words corresponds to the minimum difference between their output positions during the previous session (see Kahana and Howard, 2005). (B) Subjective clustering score as a function of session number in Experiment 1 (see Lohnas, in press) (C) Subjective lag-CRP curve for Experiment 2. The lag between two words corresponds to the minimum difference between their output positions during the previous session. (D) Subjective clustering score as a function of session number in Experiment 2. Error bars represent 95% confidence intervals. Note that in both experiments, the cumulative free recall period based on which subjective lags are computed occurs one session prior to the cumulative free recall period analyzed, hence the arrows in the x-axis of graphs B and D. Error bars in panels B and D used the Loftus-Masson (1994) adjustment for between-subject variability.

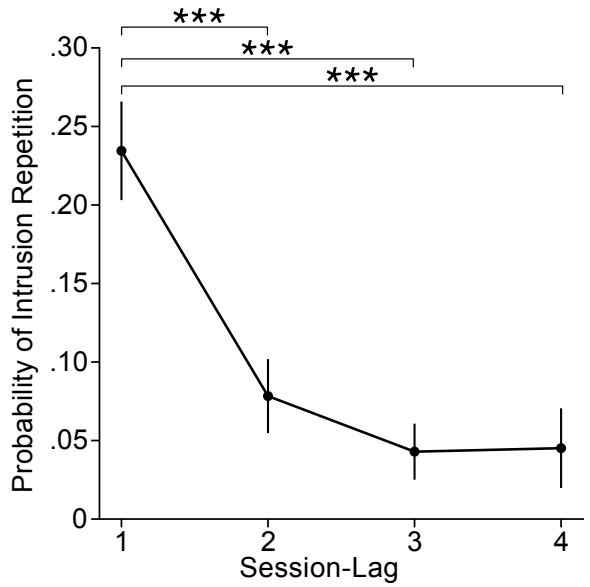

Figure 5

*The probability of repeating an intrusion for the first time as a function of the number of sessions elapsed since its first occurrence in Experiment 1. This probability was computed by counting the number of times intrusions were repeated after a given session-lag and dividing this number by the number of times intrusions could have been repeated at this lag. Post hoc pair-wise t-tests were conducted across all session-lags. *** indicates* $p < 0.0001$.