The Role of Episodic Context in Retrieval Practice Effects

Joshua W. Whiffen and Jeffrey D. Karpicke*

Purdue University

*Corresponding author. Email: karpicke@purdue.edu

Journal of Experimental Psychology: Learning, Memory, and Cognition, in press

Abstract

The episodic context account of retrieval-based learning proposes that retrieval enhances subsequent retention because people must think back to and reinstate a prior learning context. Three experiments directly tested this central assumption of the context account. Subjects studied word lists and then either restudied the words under intentional learning conditions or made list discrimination judgments by indicating which list each word had occurred in originally. Subjects in both conditions experienced all items for the same amount of time, but subjects in the list discrimination condition were required to retrieve details about the original episodic context in which the words had occurred. Making initial list discrimination judgments consistently enhanced subsequent free recall relative to restudying the words. Analyses of recall organization and retrieval strategies on the final test showed that retrieval practice enhanced temporal organization during final recall. Semantic encoding tasks also enhanced retention relative to restudying but did so by promoting semantic organization and semantically-based retrieval strategies during final recall. The results support the episodic context account of retrieval-based learning.

A wealth of recent research has examined the effects of retrieval practice on learning. When people retrieve items on an initial test, the act of initial retrieval enhances subsequent retention. Thus, the act of retrieval alters memory, making retrieved items more retrievable in the future. Retrieval practice effects are robust and have been explored with a variety of materials in a range of settings (for recent reviews, see Nunes & Karpicke, 2015; Rowland, 2014). However, there is still considerable room for progress in understanding the mechanisms of retrieval-based learning.

One recent theory of retrieval-based learning is the episodic context account (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014), which explains retrieval practice effects on the basis of four central assumptions. First, people encode information about items and the temporal/episodic context in which those items occurred (Howard & Kahana, 2002). Second, during retrieval, people attempt to reinstate the episodic context associated with an item as part of a memory search process (Lehman & Malmberg, 2013). Third, when an item is successfully retrieved, the context representation associated with that item is updated to include features of the original study context and features of the present test context. Finally, when people attempt to retrieve items again on a later test, the updated context representations aid in recovery of those items, and memory performance is improved.

The context theory can account for several key findings in the retrieval practice literature. For example, one consistent finding is that spaced retrieval produces better retention than does massed retrieval (Roediger & Karpicke, 2011). The context account proposes that temporal context will have changed more during a spaced repetition than during a massed one, so spaced retrieval may require a greater degree of context

reinstatement relative to massed retrieval. Spaced retrieval may also yield updated context representations that are more distinctive than those produced by massed retrieval (Karpicke et al., 2014). The context account also helps explain the positive effects of "effortful" initial retrieval tasks. Specifically, free recall tests tend to produce larger retrieval practice effects than do recognition tests (Glover, 1989); practicing retrieval with weakly associated cues produces larger effects relative to practicing retrieval with strong associates (Carpenter, 2009); and initial recall with only the first letter of a target as a cue produces larger retrieval practice effects than does initial recall with three letters of the target (Carpenter & DeLosh, 2006). In all cases, the conditions that produce larger retrieval practice effects (freely recalling, recalling with weak cues, and recalling with fewer letter cues) are ones that require learners to engage in greater degrees of context reinstatement during initial retrieval.

The episodic context account also helps explain the role of retrieval mode in retrieval practice effects. Retrieval mode refers to the cognitive state in which people intentionally think back to a particular place and time when an event occurred (Tulving, 1983). Experiments by Karpicke and Zaromb (2010) established the importance of retrieval mode for retrieval-based learning. In those experiments, subjects studied a list of target words (e.g., *love*) and then restudied the targets paired with related cues (e.g., *heart-love*) or saw cues and fragments of the targets (e.g., *heart-l_v_*). In one condition, subjects were told to generate words that would complete each fragment but were not told to think back to the study phase. In a second condition, subjects were placed in an episodic retrieval mode: They were told to think back to the study phase and complete the fragments with words they had studied. On final free recall and item recognition

tests, both fragment-completion conditions tended to outperform the restudy condition. Most importantly, intentionally retrieving the target words produced larger gains on the final test relative to generating the target words without recollecting the study episode (see too Pu & Tse, 2014). Thus, reinstating the original episodic context during the practice phase enhanced subsequent retention.

Although the episodic context account helps explain several key findings about retrieval practice, few studies have directly tested predictions derived from the account. The present experiments examined a central prediction: With all else held constant, if people experience items and are required to think back to an original study episode, the act of doing so should enhance subsequent retention relative to experiencing the items but not thinking back to a study episode. The present experiments accomplished this by using a list discrimination task. To implement retrieval practice, subjects were shown a list of words and indicated which list the word had occurred in during the first phase of the experiment. Prior studies have examined the effects of initial retrieval practice on later list discrimination performance (e.g., Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010; Chan & McDermott, 2007; Verkoeijen, Tabbers, & Verhage, 2011). Here, list discrimination was used as a retrieval practice task that required subjects to think back to and reinstate the original episodic context.

The list discrimination task used in the present experiments circumvents a methodological problem that often exists in retrieval practice research. In many experiments, while subjects in restudy conditions re-experience the entire set of items, subjects in retrieval practice conditions re-experience only the items they are able to recall. Thus, re-exposure to items is not equated in restudy and retrieval practice

conditions. For example, in Karpicke and Zaromb's (2010) experiments, subjects recalled approximately 70-75% of the target words during initial retrieval practice, whereas they re-experienced 100% of the targets in the restudy condition (see Karpicke et al., 2014, for further discussion of this issue). In the present experiments, subjects in all conditions re-experienced all items for the same amount of time. The only difference between the restudy and retrieval practice conditions was whether subjects were told to restudy the words or whether they were required to recollect the study episode by making list discrimination judgments.

The three experiments reported here used the same general procedure. First, subjects studied two short lists of words. Next, they were re-presented with the words from both lists mixed together. In a restudy condition, subjects were only told to restudy the words, whereas in a list discrimination condition, subjects indicated whether the words occurred in list 1 or 2. The relative effects of restudying or making list discrimination judgments were assessed on a final free recall test. The general prediction was that making list discrimination judgments would enhance final recall relative to restudying, because the list discrimination task required subjects to think back to the study episode and recollect information about the temporal occurrence of items.

Experiments 2 and 3 examined the effects of initial list discrimination on subsequent recall and also included semantic encoding conditions in which subjects made pleasantness ratings or category judgments, respectively, when they restudied the words. On the basis of vast prior research, elaborative encoding was expected to enhance recall relative to restudying. However, patterns of final recall were expected to

differ in the list discrimination and elaborative study conditions, reflecting differences in organizational output strategies used during final recall.

The episodic context account predicts that retrieval practice should produce patterns of recall output that differ from those in restudy and elaborative encoding conditions. Specifically, if context representations are updated during retrieval practice and subjects use context to guide retrieval during subsequent recall, then patterns of final recall output should show greater organization around temporal dimensions after subjects have practiced retrieval relative to when they restudied or made semantic judgments. The present experiments explored several aspects of organization and memory search dynamics during free recall. Measures of clustering were used to assess the extent to which recall was organized around the original study order. Measures of temporal and semantic factors, following Sederberg, Miller, Howard, and Kahana (2010), examined the extent to which item-to-item transitions during free recall followed the original temporal order of words or the semantic relatedness of words, respectively. Finally, an additional analysis examined the dynamics of how people searched memory during final recall, based on the idea that people forage through memory representations in ways that are similar to how animals forage in physical spaces (see Hills, Jones, & Todd, 2012; Hills, Todd, & Jones, 2015).

Experiment 1

The purpose of Experiment 1 was to test two predictions based on the episodic context account. First, making temporal judgments about when words occurred in a study list should enhance retention relative to restudying the words. In Experiment 1, subjects studied a list of words, restudied or made list discrimination judgments about

the words, then took a final free recall test. The subjects in both conditions re-experienced the words, but those in the list discrimination condition were required to think back to the original study episode and remember when the word had occurred. The effects of restudying the words or making temporal judgments were assessed on a final free recall test. The second prediction was that final recall would exhibit greater organization around the original temporal order of the items in the list discrimination condition relative to the restudy condition, because retrieval practice in the list discrimination condition would result in the reinstatement and subsequent updating of context. Analyses of temporal clustering, temporal and semantic factors, and foraging patterns during final recall were carried out to examine this prediction.

Method

Subjects. Sixty Purdue University undergraduates participated in Experiment 1 in exchange for course credit.

Materials. Thirty-six medium frequency, medium concreteness words were selected from the Clark and Paivio (2004) norms. The words were divided into six lists of six words. The lists were then paired to form three study blocks within the learning phase (lists 1-2, lists 3-4, and lists 5-6 were study blocks 1, 2, and 3, respectively). The words within each study block were equated for concreteness, imagery, and frequency, and the order of the study blocks was counterbalanced across subjects.

Design. Experiment 1 used a between-subjects design. There were two conditions, list discrimination and restudy, and 30 subjects were assigned to each condition.

Procedure. The subjects were tested in small groups of one to four people. At the beginning of the experiment, subjects were told that they would study several short lists of words and that their memory for the words would be tested at the end of the experiment. The study phase consisted of three study blocks. Within each study block, subjects studied a list of six words, performed a brief distracter task, studied a second list of six words, performed the distracter task again, and then re-experienced the 12 words in either a restudy or list discrimination task. In study periods, words were presented on a computer screen one at a time at a 3-s rate with a 500-ms interstimulus interval. In the distracter task, subjects spent 30 s solving one- or two-digit addition problems. The problems were shown one at a time on the computer, and subjects typed their answers and pressed "Enter" to advance to the next problem. After studying two lists, subjects were shown the 12 words from both lists mixed together, one at a time at a 3-s rate with a 500-ms interstimulus interval. At this point the critical manipulation occurred. In the restudy condition, subjects were instructed to restudy the list of words. In the list discrimination condition, subjects were told that they had 3 seconds to indicate whether each word was from list 1 or list 2 by clicking one of two buttons (labeled "List 1" and "List 2") shown on the computer screen. The words remained on the screen for 3 s regardless of when subjects made their responses, and the computer program automatically advanced to the next word after 3 s even if a response had not been made. Thus, in both conditions, subjects re-experienced all 12 words for the same amount of time; the difference was that one group restudied the words, while the other group was required to think back to the earlier part of the experiment and decide whether each word occurred in the first or second list. After completing the restudy or

list discrimination task, subjects completed another 30 s of the distracter task and then advanced to the next part of the experiment. This procedure, wherein subjects studied two lists and then either restudied or made list discrimination judgments, was repeated for the other two study blocks (lists 3-4 and lists 5-6), for a total of three study blocks in the learning phase.

At the end of the learning phase, subjects completed an additional 1 min of the distracter task and then took a final free recall test. On the final test, subjects were given 5 min to recall as many words as possible from the learning phase, in any order. Subjects typed their responses into a response box on the computer. They were instructed to press the "Enter" key after they had typed each response, which added that response to a list of their responses displayed on the computer screen. At the end of the experiment the subjects were debriefed and thanked for their participation.

Results

List Discrimination Performance. Overall, subjects entered responses on 99% of trials (in total, there were 1080 trials (30 subjects X 36 trials per subject), and 1065 responses were recorded). The mean proportion correct on the list discrimination task was .86. Response times were measured as the time between the onset of the word and the subject's mouse click. The average response time for correct responses was 1.6 s. Table 1 shows the mean proportion correct and mean response times across study blocks in all three experiments. In Experiment 1, list discrimination performance did not change much across study blocks, $F(2, 58) = 2.45$, $p = .10$, $\eta^2 = 0.08$, and response times tended to become slightly faster across study blocks, $F(2, 58) = 3.11$, $p = .06$, $\eta^2 = 0.10$.

Final Free Recall. The key results of Experiment 1 are the proportions of words recalled on the final free recall test, shown in the left panel of Figure 1. Subjects in the list discrimination condition recalled more items on the final test than did subjects in the restudy group (.48 vs. .38), $t(58) = 2.41$, $d = 0.62$, 95% CI [0.10, 1.14]. Thus, making a list discrimination judgment, which required people to think back to and retrieve the original temporal context in which a word occurred, produced a 10% final recall advantage relative to restudying.

Table 2 shows an analysis of the relationship between initial list discrimination performance and final free recall. Following Tulving's (1964) convention for examining the fate of individual items across two tests, C_1 refers to items correctly identified on the initial list discrimination test and N_1 refers to items that were not correct on the initial list discrimination test. C_2 refers to items recalled on the final free recall test and N_2 refers to items not recalled on the final test (see also Karpicke & Zaromb, 2010). This analysis is correlational and subject to item-selection effects. Nevertheless, the results indicate that when items were not correctly identified on the list discrimination test (N_1), it was unlikely that those items would then be recalled on the final recall test (the joint probability was .05 in Experiment 1). When items were correctly identified on the list discrimination test (C_1), they were much more likely to be recalled on the final recall test (.41 in Experiment 1).

Temporal Clustering During Final Recall*.* Clustering was measured with adjusted ratio of clustering (ARC) scores (Roenker, Thompson, & Brown, 1971). ARC scores range from -1 to 1, where 0 represents chance clustering and 1 represents perfect clustering around a dimension (negative scores are considered uninterpretable;

Murphy & Puff, 1982). ARC scores are typically calculated to measure the extent to which a person's recall output is organized around semantic (e.g., taxonomic) categories. Here, ARC scores were used to assess how well free recall was organized around study block (1, 2, or 3). The right panel in Figure 1 shows the mean temporal clustering scores. Subjects in the list discrimination condition had higher temporal clustering scores than did subjects in the restudy condition (.38 vs. .25), $t(58) = 1.77$, $d = 0.46$ [-0.06, 0.97]. The temporal clustering scores indicate that subjects in the list discrimination condition organized their recall around the original study order more than subjects in the restudy condition did, which supports the idea that these subjects used episodic context information to guide their output during recall.

Temporal and Semantic Factors During Final Recall. To further substantiate this interpretation, measures of temporal and semantic factors were calculated following the methods proposed by Sederberg et al. (2010). Temporal factors reflect the degree to which transitions during free recall output followed the original temporal order in which words were studied. Semantic factors reflect the degree to which transitions during recall followed the semantic relatedness of the words, which was defined as the similarity scores for each pair of words in the study list based on Latent Semantic Analysis (Landauer & Dumais, 1997). Briefly, temporal and semantic factors for each recall protocol were calculated in the following way. For each transition, all possible transitions were ranked according to temporal proximity or semantic relatedness for temporal or semantic factors, respectively. The rank of the actual transition relative to all other possible transitions was determined, and each transition received a score from 0 to 1, with 1 representing the closest transition and 0 representing the farthest. The

average of the scores represented the temporal or semantic factor for each protocol (see Sederberg et al., 2010, for details). Therefore, temporal and semantic factors range from 0 to 1, where factors closer to 1 indicate that subjects transitioned to the most temporally or semantically proximal words during recall, and factors closer to 0 indicate that subjects transitioned to the least temporally or semantically proximal words during recall.

Subjects in the list discrimination condition showed larger temporal factors than did subjects in the restudy condition (.71 vs. .61), $t(58) = 3.64$, $d = 0.94$ [0.40, 1.47], consistent with the temporal clustering analysis carried out with ARC scores. In contrast, there was essentially no difference in the semantic factors in the list discrimination and restudy conditions (.54 vs. .55), $t(58) = 0.18$, $d = 0.05$ [-0.46, 0.55].

Foraging Patterns During Final Recall. The final analysis examined the dynamics of how people searched memory during final recall, based on the idea that people search memory in ways that are similar to how animals forage in physical environments (Hills et al., 2015). Specifically, people search memory by visiting sets of items, referred to here as "patches," and spend time recovering items from one patch before switching and searching a different patch. The analyses of temporal clustering and temporal factor suggested that retrieval practice produced memory structures that were organized around temporally-defined patches (study blocks 1, 2, and 3). The foraging analysis explored this further by examining transitions to and from each temporal patch during free recall. The onset of a temporal patch visit occurred when a subject recalled an item from a study block that differed from the study block of the item recalled immediately before it, and the end of a patch visit was defined as the onset of

recall from another patch. Subjects with well-defined structures, created by practicing retrieval during learning, may engage in more efficient searches than do subjects with memory structures that are not as well defined. In particular, they may visit temporal patches fewer times, recover more items per visit, and spend more time searching per visit.

Overall, the mean number of patch visits did not differ between the list discrimination and restudy conditions (7.30 vs. 7.37), $t(58) = 0.09$, $d = 0.02$ [-0.48, 0.53]. However, subjects recovered more items per visit in the list discrimination condition than they did in the restudy condition (2.61 vs. 1.95), $t(58) = 2.53$, $d = 0.65$ [0.13, 1.17]. Subjects also spent more time searching during each patch visit in the list discrimination condition than they did in the restudy condition (26.5 s vs. 18.8 s), $t(58) = 2.28$, $d = 0.59$ [0.07, 1.10]. Table 3 shows the mean number of items recovered as a function of visit number, and Table 4 shows the mean time spent searching as a function of visit number. These data illustrate that differences in number of items recovered and search times in the list discrimination and restudy conditions were pronounced during the first few visits, early in the recall period, and became less pronounced during later recall.

Discussion

Experiment 1 provided evidence consistent with the episodic context account of retrieval practice. Subjects were required to make a list discrimination judgment as a retrieval practice activity. The task required subjects to think back to the study episode and determine when each item had occurred in the study phase. All items were re-presented to subjects in both conditions for the same amount of time; the only difference between conditions was whether subjects made a judgment about the

previous occurrence of the items. Experiment 1 showed that the act of making list discrimination judgments produced a retrieval practice effect, enhancing subsequent recall relative to restudying. In addition, clustering analyses indicated that subjects in the list discrimination condition used the original study order as a strategy to guide recall output, which further supports the context account of retrieval practice. Experiment 2 was aimed at expanding upon these findings.

Experiment 2

The goals of Experiment 2 were to replicate the main findings from Experiment 1 and to compare the effects of making list discrimination judgments to the effects of making semantic judgments. The procedure was identical to the one used in Experiment 1 with the addition of a pleasantness rating condition. Rating the pleasantness of words is a widely used semantic encoding task that, unlike list discrimination, does not require subjects to engage in episodic remembering. The effects of the three learning conditions were assessed on a final free recall test and with analyses of the organization of recall around episodic and semantic dimensions.

Method

Subjects. One hundred and twenty Purdue University undergraduates participated in Experiment 2 in exchange for course credit. None of the subjects had participated in Experiment 1. The number of subjects in Experiment 2 was larger than the number in Experiment 1 to improve power and the precision of effect size estimates.

Materials. A new set of 36 medium frequency, medium concreteness words was selected from the Clark and Paivio (2004) norms. As in Experiment 1, the words were divided into six lists of six words and then paired to form three study blocks within the

learning phase. The words within each study block were equated for concreteness, imagery, word frequency, and pleasantness as determined by the ratings reported in Clark and Paivio (2004). The order of the study blocks was counterbalanced across subjects. Pleasantness was equated such that each list pair had the same number of words from each normed pleasantness rating (e.g., two words with a normative pleasantness rating of 1, two with a rating of 2, etc.).

Design. Experiment 2 used a between-subjects design. There were three conditions: list discrimination, restudy, and pleasantness. Forty subjects were assigned to each condition.

Procedure. The procedure was identical to the one used in Experiment 1, with the addition of the pleasantness condition. The procedure involved three phases: Subjects studied a list of words, then restudied or made judgments about the words, and then took a final free recall test. The procedures used in the restudy and list discrimination conditions were identical to those used in Experiment 1. In the pleasantness condition, when subjects were re-exposed to the list of words, they rated the pleasantness of each word on a scale from 1 (very pleasant) to 7 (very unpleasant) by clicking one of the seven corresponding radio buttons displayed below the word. The words remained on the screen for 3 s regardless of when subjects made their responses, and the computer program automatically advanced to the next word after 3 s even if a response had not been made. In all conditions, subjects re-experienced the words for the same amount of time; the difference was whether subjects restudied the words, rated the pleasantness of the words, or made a list discrimination decision about the words by thinking back to the prior study episode.

Results

List Discrimination Performance. Subjects entered responses on 96% of trials (in total, there were 1440 trials (40 subjects X 36 trials per subject), and 1389 responses were recorded). The mean proportion correct on the list discrimination task was .85, and the mean response time for correct responses was 1.7 s. As shown in Table 1, there was little change in list discrimination performance across study blocks, $F(2, 78) = 1.21$, $\eta^2 = 0.08$, and, contrary to the results of Experiment 1, response times did not differ much across study blocks, $F(2, 78) = 0.09$, $\eta^2 = 0.00$.

Pleasantness Rating Performance. In the pleasantness condition, subjects entered responses on 95% of trials (1370 responses out of a total of 1440 trials). The mean response time was 2.0 s.

Final Free Recall. The left panel of Figure 2 shows the proportion of words recalled on the final free recall test. As in Experiment 1, subjects in the list discrimination condition recalled more words than did subjects in the restudy condition (.43 vs. .31), $t(78) = 3.27$, $d = 0.73$ [0.28, 1.18]. Subjects in the pleasantness condition also outperformed subjects in the restudy condition (.41 vs. .31), $t(78) = 3.51$, $d = 0.78$ [0.33, 1.23]. There was little difference in recall between the list discrimination and pleasantness conditions, $t(78) = 0.32$, $d = 0.07$ [-0.37, 0.51]. Making list discrimination and pleasantness judgments enhanced final recall relative to restudying the words.

The middle row in Table 2 shows the relationship between initial list discrimination performance and final free recall in Experiment 2. When items were not correctly identified on the list discrimination test, it was unlikely that those items were recalled on the final recall test (the joint probability was .04). When items were correctly

identified on the list discrimination test, they were much more likely to be recalled on the final recall test (.38).

Temporal Clustering During Final Recall. The right panel of Figure 2 shows temporal clustering scores, which were ARC scores that assessed the extent to which recall was organized around study block. Temporal clustering scores were higher in the list discrimination condition than they were in the restudy condition (.38 vs. .24), $t(78) = 2.16$, $d = 0.48$ [0.04, 0.93] and in the pleasantness condition (.38 vs. .16), $t(78) = 3.89$, $d = 0.87$ [0.41, 1.33]; for pleasantness vs. restudy, $t(78) = 1.45$, $d = 0.32$ [-0.12, 0.76]. When subjects made list discrimination judgments during the learning phase, they subsequently used temporal context information to guide free recall, consistent with the episodic context account.

An additional analysis examined the extent to which recall was organized around pleasantness ratings. Pleasantness clustering scores were calculated as ARC scores with normative pleasantness ratings (from Clark & Paivio, 2004) as the organizing dimension. The highest pleasantness clustering scores were observed in the restudy condition and were similar to those in the pleasantness rating condition (.17 vs. .14), $t(78) = 0.49$, $d = 0.11$ [-0.33, 0.55]. Pleasantness clustering scores were slightly lower in the list discrimination condition than they were in the restudy condition (.07 vs. .17), $t(78) = 1.67$, $d = 0.37$ [-0.07, 0.81], and in the pleasantness rating condition (.07 vs. .14), $t(78) = 1.33$, $d = 0.30$ [-0.14, 0.74]. In general, however, pleasantness clustering scores were similar across all conditions. Thus, normative pleasantness did not produce large influences on the organization of final recall.

Temporal and Semantic Factors During Final Recall. The analyses of temporal and sematic factors during final recall provided further evidence that subjects in the list discrimination and pleasantness rating conditions used different strategies during the final recall task. Subjects in the list discrimination condition had higher temporal factors relative to subjects in the restudy condition (.67 vs. .59), $t(78) = 3.16$, $d = 0.71$ [0.25, 1.16] and subjects in the pleasantness condition (.67 vs. .55), $t(78) = 6.03$, $d = 1.35$ [0.86, 1.83]. Temporal factors were slightly higher in the restudy condition than they were in the pleasantness condition, $t(78) = 1.69$, $d = 0.38$ [-0.07, 0.82]. In contrast, semantic factors were similar across conditions. As in Experiment 1, semantic factors were similar in the list discrimination and restudy conditions (.51 vs. .53), $t(78) = 0.51$, $d = 0.11$ [-0.32, 0.55]. Likewise, the factors were similar in the pleasantness and restudy conditions (.50 vs. .53), $t(78) = 1.24$, $d = 0.28$ [-0.16, 0.72]; for list discrimination vs. pleasantness, $t(78) = 0.90$, $d = 0.20$ [-0.24, 0.64].

Foraging Patterns During Final Recall. The mean number of temporal patch visits was greater in the list discrimination condition than in the restudy condition (6.70 vs. 5.78), $t(78) = 1.62$, $d = 0.36$ [-0.08, 0.80], and the number of items recovered per visit was greater in the list discrimination condition than in the restudy condition (2.57 vs. 2.15), $t(78) = 1.74$, $d = 0.39$ [-0.05, 0.83]. Search times during each patch visit were slightly longer in the list discrimination condition than in the restudy condition (24.6 s vs. 22.7 s), $t(78) = 0.58$, $d = 0.13$ [-0.31, 0.57]. In the pleasantness condition, the number of patch visits was greater than it was in the list discrimination condition (8.90 vs. 6.70), $t(78) = 3.61$, $d = 0.81$ [0.50, 1.26], and in the restudy condition (8.90 vs. 5.78), $t(78) = 5.12$, $d = 1.14$ [0.67, 1.62]. Subjects in the pleasantness condition also recovered fewer

items per visit relative to subjects in the list discrimination condition (1.79 vs. 2.57), $t(78) = 3.58$, $d = 0.80$ [0.34, 1.25], and subjects in the restudy condition (1.79 vs. 2.15), $t(78) = 2.17$, $d = 0.62$ [0.17, 1.07]. Finally, search times during each patch visit were slightly shorter in the pleasantness condition than they were in the list discrimination condition (21.4 s vs. 24.6 s), $t(78) = 1.10$, $d = 0.25$ [-0.19, 0.69]. Search times were similar in the pleasantness and restudy conditions (21.4 s vs. 22.7 s), $t(78) = 0.39$, $d = 0.09$ [-0.35, 0.53]. Tables 3 and 4 shows that the largest differences in number of items recovered and search times, respectively, occurred during the first few visits, early in the recall period.

Discussion

Experiment 2 replicated the key findings from Experiment 1. Making list discrimination judgments enhanced final recall and increased the temporal organization of recall relative to restudying. Making semantic judgments (pleasantness ratings) also enhanced subsequent recall but did not increase the degree of temporal organization in final recall, as evidenced by the analyses of temporal clustering, temporal factors, and foraging patterns during final recall. The results provide further support for the idea that reinstating the episodic context during initial learning improved subsequent recall and promoted greater temporal organization on the final test.

Experiment 3

Experiment 3 provided an additional examination of the effects of retrieval practice on temporal and semantic organizational factors in recall. The procedure in Experiment 3 followed the procedure used in the previous experiments, except that subjects studied categorized lists of words rather than unrelated lists. The experiment

involved three conditions. In addition to restudy and list discrimination conditions, which were identical to those used in previous experiments, Experiment 3 included a category judgment condition in which subjects identified the taxonomic categories of the words. The category judgment task oriented subjects to semantic attributes of the words and, like the pleasantness rating task in Experiment 2, did not require subjects to think back to the study episode. Whereas pleasantness ratings are thought to promote retention by emphasizing the distinctiveness of items, category judgments require subjects to process how items are related to an organizational scheme. The analyses conducted in the previous two experiments were also conducted in Experiment 3 with the addition of analyses of clustering around semantic categories during free recall (traditional ARC scores) and memory foraging patterns based on semantic categories.

Method

Subjects*.* One hundred and twenty Purdue University undergraduates participated in this experiment in exchange for course credit. None of the subjects had participated in Experiments 1 or 2.

Materials. Thirty-six words were selected from the Van Overschelde, Rawson, and Dunlosky (2004) norms. The most frequent six exemplars were selected from six taxonomic categories (*animals, fruits, body parts, clothing, instruments*, and *insects*). As in the previous experiments, the words were assigned to six lists of six words. One word from each category was assigned to each list.

Design. Experiment 3 used a between-subjects design. There were three conditions: list discrimination, restudy, and category judgment. Forty subjects were assigned to each condition.

Procedure. The procedure was identical to the one used in Experiment 1, with the addition of the category judgment condition. Subjects studied a list of words, then restudied or made judgments about the words, and then took a final free recall test. The restudy and list discrimination conditions were identical to those in the previous experiments. In the category judgment condition, subjects saw each word and two category alternatives (e.g., for the word *banana*, subjects might see *fruits* and *animals* as alternatives). Subjects indicated which category the word belonged to by clicking a button associated with the alternative. The words remained on the screen for 3 s regardless of when subjects made their responses, so that subjects in all conditions re-experienced the words for the same duration.

Results

List Discrimination Performance. Subjects entered responses on 97% of trials (1396 responses on 1440 trials). The mean proportion correct on the list discrimination task was .81, and the average response time for correct responses was 1.7 s. As shown in Table 1, there were differences in list discrimination performance across blocks, $F(2, 78) = 5.00$, $\eta^2 = 0.11$, and response times tended to become faster across blocks, $F(2, 78) = 3.31$, $\eta^2 = 0.08$.

Category Judgment Performance. In the category judgment condition, subjects entered responses on 98% of trials (1416 responses on 1440 trials). The mean proportion correct was .99, and the mean response time for correct responses was 1.7 s.

Final Free Recall. Figure 3 shows the proportion of words recalled on the final free recall test. As in Experiments 1 and 2, subjects in the list discrimination condition group recalled more words than subjects in the restudy condition (.55 vs. .49), $t(78) =$

2.12, *d* = 0.47 [0.02, 0.92]. Subjects in the category judgment condition slightly outperformed subjects in the restudy condition by a small amount (.53 vs. .49), *t*(78) = 1.41, *d* = 0.25 [-0.19, 0.69]. There was little difference between the list discrimination and category sorting conditions, *t*(78) = 0.54, *d* = 0.12 [-0.32, 0.56].

The bottom row in Table 2 shows the relationship between initial list discrimination performance and final free recall in Experiment 3. When items were not correctly identified on the list discrimination test, those items were not likely to be recalled on the final recall test (the joint probability was .08). When items were correctly identified on the list discrimination test, they were much more likely to be recalled on the final recall test (.45).

Temporal and Semantic Clustering During Final Recall. The middle panel of Figure 3 shows temporal clustering scores, calculated as they were in previous experiments. Temporal clustering scores were higher in the list discrimination condition than they were in the restudy condition (.25 vs. .15), *t*(78) = 1.84, *d* = 0.41 [-0.03, 0.85], and in the category judgment condition (.25 vs. .08), *t*(78) = 3.71, *d* = 0.83 [0.37, 1.28]; for category judgment vs. restudy, *t*(78) = 1.56, *d* = 0.35 [-0.09, 0.79]. The right panel of Figure 3 shows semantic clustering scores, which were ARC scores with taxonomic category as the organizing dimension. The category judgment task produced the highest semantic clustering scores, higher than scores in the list discrimination condition (.41 vs. .21), *t*(78) = 3.67, *d* = 0.82 [0.36, 1.27], and slightly higher than those in the restudy condition (.41 vs. .34), *t*(78) = 1.09, *d* = 0.24 [-0.20, 0.68]; for list discrimination vs. restudy, *t*(78) = 2.33, *d* = 0.52 [0.07, 0.97]. Thus, the pattern of semantic clustering scores was the opposite of the pattern of temporal clustering scores.

Temporal and Semantic Factors During Final Recall. The list discrimination condition produced higher temporal factors relative to the restudy condition (.60 vs .55), $t(78) = 2.09$, $d = 0.47$ [0.02, 0.91], and the category judgment condition (.60 vs .53), $t(78) = 2.63$, $d = 0.59$ [0.14, 1.03]. There was little difference between the temporal factors in the restudy and category judgment conditions, $t(78) = 0.07$, $d = 0.02$ [-0.42, 0.45]. However, the semantic factors showed a different pattern of results. Semantic factors were slightly higher in the category judgment condition relative to the restudy condition (.64 vs. .61), $t(78) = 1.34$, $d = 0.30$ [-0.14, 0.74] and the list discrimination condition (.64 vs. .58), $t(78) = 2.12$, $d = 0.47$ [0.03, 0.92]; for list discrimination vs. restudy, $t(78) = 1.52$, $d = 0.34$ [-0.10, 0.78]. Overall, the patterns of temporal and semantic factors across conditions matched the patterns of temporal and semantic clustering in final recall.

Foraging Patterns During Final Recall. In Experiment 3, search patterns during final recall could have relied on temporal patches (study blocks 1, 2, and 3) or semantic patches (taxonomic categories). Thus, both possible ways of searching memory were analyzed.

When examining foraging based on a temporal search strategy, the mean number of temporal patch visits was numerically smaller for list discrimination relative to restudy (10.10 vs. 11.20), $t(78) = 1.14$, $d = 0.25$ [-0.18, 0.69], and relative to category judgment (10.10 vs. 12.30), $t(78) = 2.56$, $d = 0.57$ [0.12, 1.02]; for restudy vs. category, $t(78) = 1.07$, $d = 0.24$ [-0.20, 0.68]. Also, the mean number of items recovered per visit was greater for list discrimination compared to restudy (2.07 vs. 1.66), $t(78) = 2.78$, $d = 0.62$ [0.17, 1.07], and category judgment (2.07 vs. 1.59), $t(78) = 3.45$, $d = 0.77$ [0.31,

1.22]; for restudy vs. category, $t(78) = 0.99$, $d = 0.22$ [-0.21, 0.66]. Subjects spent more time per visit in list discrimination compared to restudy (18.3 vs. 14.3 s), $t(78) = 2.20$, d = 0.49 [0.05, 0.94], and category judgment (18.3 vs. 14.9), $t(78) = 1.95$, d = 0.44 [0.00, 0.88]; for restudy vs. category, $t(78) = 0.39$, d = 0.09 [-0.35, 0.68]. As in the previous experiments, Tables 3 and 4 show that in Experiment 3, the largest differences in number of items recovered and search times, respectively, occurred during the first few visits, early in the recall period.

The foraging analysis with semantic category as "patch" was conducted the same way as the analysis of temporal patches except the patches were the taxonomic categories used in the experiment (i.e., fruit, clothing, animals, instruments, body parts, and insects). The mean number of semantic patch visits was greater in the list discrimination condition compared to restudy (14.63 vs. 12.08), $t(78) = 2.30$, $d = 0.51$ [0.07, 0.96], and category judgment (14.63 vs. 11.9), $t(78) = 2.69$, $d = 0.60$ [0.15, 1.05]; for restudy vs. category, $t(78) = 0.17$, $d = 0.04$ [-0.40, 0.48]. However, the mean number of items recovered per visit was smaller for list discrimination compared to restudy (1.38 vs. 1.53), $t(78) = 1.74$, $d = 0.39$ [-0.05, 0.83], and category judgment (1.38 vs. 1.67), $t(78) = 3.07$, $d = 0.69$ [0.23, 1.14]; for restudy vs. category, $t(78) = 1.32$, $d = 0.30$ [-0.14, 0.74]. Subjects spent slightly more time per visit in the list discrimination condition relative to restudy (15.5 vs. 13.7), $t(78) = 1.23$, d = 0.28 [-0.17, 0.71], and relative to category judgment (15.5 vs. 15.14), $t(78) = 0.20$, d = 0.04 [-0.39, 0.48]; for restudy vs. category, $t(78) = 0.89$, d = 0.19 [-0.24, 0.64].

Discussion

Experiment 3 replicated the key findings from the previous experiments. Making list discrimination judgments led to enhanced recall and temporally organized output relative to restudy. However, this experiment was also able to examine semantic organization in recall and found that making list discrimination judgments led to a temporal output strategy, but making sorting words into categories (semantic judgments) led to a semantically based output strategy. Further, analyses of search patterns replicated the previous experiments, but also showed that the category judgment and restudy conditions were searched memory based on semantically-defined patches of information while the list discrimination condition searched based on temporally-defined patches. This dissociation in how recall was organized further supports the episodic context account that reinstatement of temporal context allows subjects to use temporal information to guide output on the criterial test.

General Discussion

The purpose of this project was to evaluate the core assumptions of the episodic context account of retrieval-based learning. The account proposes that when people engage in retrieval, they attempt to reinstate the context of a prior learning episode. When retrieval is successful, the context representation associated with retrieved items is updated to include features of the retrieved context and features of the present context. Consequently, when people attempt to retrieve items again in the future, the updated context representations facilitate retrieval of those items, and memory performance is improved relative to situations in which people had not practiced retrieval.

The present experiments examined predictions that follow from the episodic context account. One prediction was that when subjects experience an item, thinking back to a prior occurrence of that item should enhance subsequent retention relative to conditions in which, with all else held constant, people do not think back to a prior occurrence. The present experiments manipulated the retrieval of occurrence information with a list discrimination task, which required subjects to make explicit judgments about when items had occurred in a previous study episode. In all three experiments, initial list discrimination enhanced final recall relative to restudying items under intentional learning instructions. It is important to emphasize that subjects re-experienced the entire list in both conditions. The only difference between conditions was that subjects in the list discrimination condition were asked to think back to the prior occurrence of the words while subjects in the restudy condition were not. To assess the overall results across the three experiments in this report, overall effect sizes comparing the list discrimination condition to the restudy condition were calculated using weighted effect sizes and a fixed effect meta-analysis model. The overall effect of retrieval practice in the list discrimination condition relative to restudying was d = 0.61 [0.34, 0.88].

A second prediction derived from the episodic context account was that initial retrieval practice would enhance the degree to which final recall was organized around the original temporal order of events. Patterns of organization during final recall were assessed in several converging ways. Relative to the restudy control condition, retrieval practice in the list discrimination condition enhanced the degree to which items were clustered around the original study order (the overall effect was d = 0.45 [0.18, 0.72]).

Measures of temporal and semantic factors (Sederberg et al., 2010) assessed the extent to which item-to-item transitions during free recall followed the original temporal order of words or the semantic relatedness of words. Retrieval practice enhanced temporal factors during final recall, d = 0.68 [0.41, 0.95], but there was little effect on semantic factors, d = 0.18 [-0.09, 0.44]. Finally, a foraging analysis (Hills et al., 2012, 2015) examined the dynamics of how people searched memory during final recall. Comparing the list discrimination and restudy conditions, there was no difference in the number of times subjects visited temporally-defined patches during memory search, d = 0.04 [-0.22, 0.31], but subjects in the list discrimination condition recovered more items per visit, d = 0.54 [0.27, 0.81], and spent more time searching per visit, d = 0.38 [0.12, 0.65], than did subjects in the restudy condition. Practicing retrieval had clear and consistent effects on search strategies during the final recall test.

Practicing retrieval in the list discrimination condition produced patterns of final recall that differed from those produced by elaborative study tasks, including rating the pleasantness of words (Experiment 2) and judging category membership (Experiment 3). Both elaborative study tasks enhanced retention relative to restudying the words, which was no surprise, since elaborative encoding has been shown to enhance retention in decades of research. However, while retrieval practice enhanced temporal organization during final recall – as assessed with temporal ARC scores, temporal factors, and foraging analyses – elaborative encoding tasks did not. For instance, the pleasantness and category judgment tasks resulted in the least amount of temporal clustering in Experiments 2 and 3 (see Figures 2 and 3), even less temporal clustering than that in the restudy control condition. Whereas final recall in the retrieval practice condition was

clearly organized around temporal dimensions, recall in the elaborative encoding conditions tended to be more closely based on semantic factors.

Previous studies have compared retrieval practice to elaborative study conditions and reasoned that if retrieval-based learning is due to elaboration, then elaborative study and retrieval practice tasks should produce the same final performance (see Karpicke & Blunt, 2011; Karpicke & Smith, 2012; Lehman et al., 2014). Those studies showed that retrieval practice and elaboration produce different final test performance, which casts doubt on the idea that the same mechanism or strategy was responsible for both effects. In Experiments 2 and 3 in the present report, there was little difference between retrieval practice and elaboration conditions (pleasantness and category sorting) on final free recall. One might be tempted to conclude that similar final test performance affirms that retrieval practice effects are due to elaboration. However, this reasoning would not be valid, because it relies on affirming the consequent. Two different tasks can produce the same level of performance via different mechanisms or strategies, and the present experiments provide a prime example. The clustering measures, temporal and semantic factors, and foraging analyses showed that retrieval practice and elaborative study tasks yielded very different patterns of final recall organization, suggesting that the effects were driven by different mechanisms in different conditions.

It is worth considering the present findings in light of alternative explanations of retrieval practice, such as the elaborative retrieval account (Carpenter, 2009, 2011). This account proposes that as people search for target items during the process of retrieval, other items that are semantically related to the retrieval cue (related words, or

mediators) become activated. This semantic elaboration assumed to occur during initial retrieval is also thought to be responsible for enhancing retention on a subsequent test (see Lehman & Karpicke, 2016). It is not readily apparent how the elaborative retrieval account might explain the present results. Making list discrimination judgments is, by definition, an episodic task, and it is not clear why any activation of semantically related words would occur when people attempt to judge the list membership of individual words. Even if list discrimination judgments did induce such semantic elaboration, it would be hard to reconcile the elaborative retrieval account with the present analyses of final recall, which show that retrieval practice produced temporally organized recall and, in some instances, reduced semantic organization (e.g., see Figure 3). Retrieval practice reliably enhanced retention in the present experiments, but that enhancement was driven by temporal factors, not semantic ones.

The episodic context account of retrieval-based learning shares some similarities with ideas that have been proposed to explain spaced repetition effects (see Karpicke et al., 2014). Specifically, a spaced repetition may enhance retention because the repetition reminds the learner of a previous occurrence (e.g., Wahlheim & Jacoby, 2013) or, similarly, because the repetition affords retrieval of a prior occurrence (an idea known as study-phase retrieval; e.g., Benjamin & Tullis, 2010). Wahlheim and Jacoby proposed that when a person is reminded of a prior occurrence, the representation of first presentation is "included" in the representation of the second presentation. Raaijmakers (2003) implemented the idea of study-phase retrieval in the SAM model (Raaijmakers & Shiffrin, 1981). In Raaijmakers's account, when a studied item is repeated, people may retrieve the trace of the prior presentation and, when this

happens, the context strength associated with that item is incremented (Raaijmakers's model incorporates additional assumptions about contextual variability; see too Delaney, Verkoeijen, & Spirgel, 2010). As discussed by Karpicke et al. (2014), these accounts of spacing effects share several features with episodic context account of retrieval practice. One difference, however, is that in studies of spaced repetition, the processes of reminding or study-phase retrieval are incidental, assumed to occur spontaneously, whereas people are explicitly prompted to think back to a prior occurrence when they practice retrieval. Most importantly, the ideas of reminding and study-phase retrieval attribute the benefits of spaced repetition to retrieval practice, which is itself a phenomenon that needs to be explained. The ideas in the episodic context account therefore add to reminding and study-phase retrieval theories by proposing mechanisms to explain how the process of retrieval enhances subsequent retention.

The present project tested the core assumptions of the episodic context account of retrieval-based learning and provided evidence supporting the account. Thinking back to a prior learning episode – an essential ingredient of retrieval practice – enhances later retention and produces fundamental changes in how learners organize subsequent recall.

References

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228-247. doi:10.1016/j.cogpsych.2010.05.004

Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory, 18*(4), 385-393. doi:10.1080/09658211003702163

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569. doi:10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547-1552. doi:10.1037/a0024140

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268-276. doi:10.3758/bf03193405

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 431-437. doi:10.1037/0278-7393.33.2.431

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments & Computers, 36*(3), 371-383. doi:10.3758/bf03195584

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 53, pp. 63-147). San Diego, CA: Elsevier Academic Press.

Glover, J. A. (1989). The 'testing' phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392-399. doi:10.1037/0022-0663.81.3.392

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431-440. doi:10.1037/a0027373

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science, 7*(3), 513-534. doi:10.1111/tops.12151

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269-299. doi:10.1006/jmps.2001.1388

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 61, pp. 237-284). San Diego, CA: Elsevier Academic Press.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227-239. doi:10.1016/j.jml.2009.11.010

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of

knowledge. *Psychological Review, 104*(2), 211-240. doi:10.1037/0033-295x.104.2.211

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/xlm0000267

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review, 120*(1), 155-189. doi:10.1037/a0030851

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1787-1794. doi:10.1037/xlm0000012

Murphy, M. D., & Puff, C. R. (1982). Free recall: Basic methodology and analysis. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 99-128). San Diego, CA: Academic Press.

Nunes, L. D., & Karpicke, J. D. (2015). Retrieval-based learning: Research at the interface between cognitive science and education. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (pp. 1-16): John Wiley & Sons, Inc.

Pu, X., & Tse, C.-S. (2014). The influence of intentional versus incidental retrieval practices on the role of recollection in test-enhanced learning. *Cognitive Processing, 15*(1), 55-64. doi:10.1007/s10339-013-0580-2

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27*(3), 431-452. doi:10.1016/s0364-0213(03)00007-7

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134. doi:10.1037/0033-295x.88.2.93

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: Essays in Honor of Robert A. Bjork* (pp. 23-48). New York: Psychology Press.

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76*(1), 45-48. doi:10.1037/h0031355

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:10.1037/a0037559

Sederberg, P. B., Miller, J. F., Howard, M. W., & Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition, 38*(6), 689-699. doi:10.3758/mc.38.6.689

Tulving, E. (1964). Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychological Review, 71*(3), 219-236. doi:10.1037/h0043186

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

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*(3), 289-335. doi:10.1016/j.jml.2003.10.003

Verkoeijen, P. P. J. L., Tabbers, H. K., & Verhage, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology, 58*(6), 490-498. doi:10.1027/1618-3169/a000117

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory & Cognition, 41*(1), 1-15. doi:10.3758/s13421-012-0246-9

Acknowledgements

This research was supported in part by grants from the National Science Foundation (DRL-1149363 and DUE-1245476) and the Institute of Education Sciences in the U.S. Department of Education (R305A110903 and R305A150546). The opinions expressed are those of the authors and do not represent the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education. We thank Nola Daley and Nick Counger for help collecting the data, Philip Grimaldi for help with computer programming, and James Nairne and Greg Francis for comments.

Table 1

Mean proportion correct and response time (in milliseconds) on the list discrimination tasks in all experiments

	Proportion Correct	Response Time
Experiment 1		
Block 1	.87 (.03)	1722 (60)
Block 2	.89 (.03)	1501 (63)
Block 3	.82 (.03)	1581 (78)
Experiment 2		
Block 1	.88 (.02)	1686 (57)
Block 2	.84 (.02)	1708 (71)
Block 3	.85 (.02)	1676 (64)
Experiment 3		
Block 1	.86 (.02)	1810 (70)
Block 2	.76 (.03)	1740 (66)
Block 3	.82 (.03)	1620 (58)

Note. Standard errors are in parentheses.

Table 2

Fates of individual items in the list discrimination conditions: Joint probabilities between initial list discrimination performance and final free recall

	C_1C_2	C_1N_2	N_1C_2	N_1N_2
Experiment 1	.41 (.03)	.44 (.03)	.05 (.03)	.10 (.03)
Experiment 2	.38 (.03)	.47 (.02)	.04 (.01)	.10 (.01)
Experiment 3	.45 (.03)	.36 (.02)	.08 (.01)	.10 (.01)

Note. Standard errors are in parentheses. C_1 = correct on the initial list discrimination task. N_1 = not correct on the initial list discrimination task. C_2 = items successfully recalled on the final free recall test. N_2 = items not recalled on the final free recall test.

Table 3

Mean number of items recovered as a function of visit for all conditions in all experiments

	Visit 1	Visit 2	Visit 3	Visit 4
Experiment 1				
List Discrimination	3.30 (.46)	2.97 (.37)	2.24 (.28)	2.25 (.29)
Restudy	2.50 (.40)	2.47 (.30)	2.03 (.25)	1.66 (.17)
Experiment 2				
List Discrimination	4.13 (.37)	2.28 (.25)	2.31 (.31)	1.94 (.24)
Restudy	2.33 (.28)	2.15 (.19)	2.54 (.28)	1.81 (.17)
Pleasantness	2.18 (.27)	1.93 (.21)	1.90 (.19)	1.77 (.15)
Experiment 3				
List Discrimination	3.48 (.49)	1.93 (.23)	2.13 (.25)	2.00 (.22)
Restudy	2.00 (.20)	1.80 (.26)	1.58 (.14)	1.56 (.15)
Category Judgment	1.90 (.17)	1.53 (.13)	1.56 (.14)	1.46 (.12)

Note. The results are only reported up to the fourth visit because not all subjects had responses for five or more visits. Standard errors are in parentheses.

Table 4

Mean time (in seconds) spent within each patch as a function of visit for all conditions in all experiments

	Visit 1	Visit 2	Visit 3	Visit 4
Experiment 1				
List Discrimination	10.7 (1.9)	12.6 (3.8)	14.2 (4.7)	24.2 (8.3)
Restudy	6.5 (0.9)	6.1 (1.1)	7.1 (2.0)	5.7 (0.9)
Experiment 2				
List Discrimination	12.4 (1.8)	10.6 (2.7)	19.9 (4.2)	27.9 (7.5)
Restudy	6.5 (0.9)	7.7 (1.6)	29.5 (9.5)	13.96 (2.8)
Pleasantness	6.8 (0.9)	5.6 (0.9)	18.0 (6.9)	15.6 (3.8)
Experiment 3				
List Discrimination	10.7 (2.1)	4.6 (0.9)	8.3 (2.4)	9.3 (2.6)
Restudy	5.6 (0.8)	4.2 (0.7)	5.1 (1.2)	5.6 (1.1)
Category Judgment	5.2 (0.6)	5.1 (1.3)	4.1 (0.5)	4.3 (1.0)

Note. The results are only reported up to the fourth visit because not all subjects had responses for five or more visits. Standard errors are in parentheses.

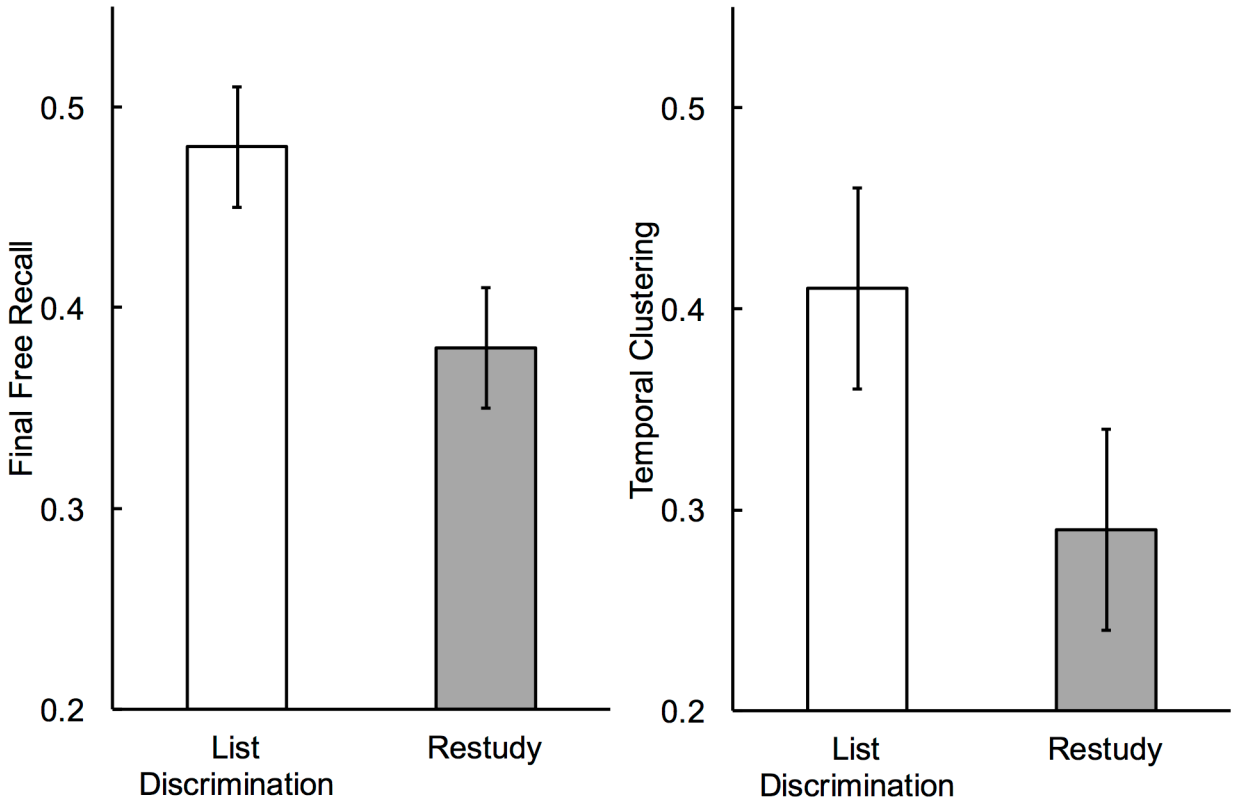

Figure 1. Proportion correct on final free recall and temporal clustering scores in Experiment 1. Error bars represent standard errors of the mean.

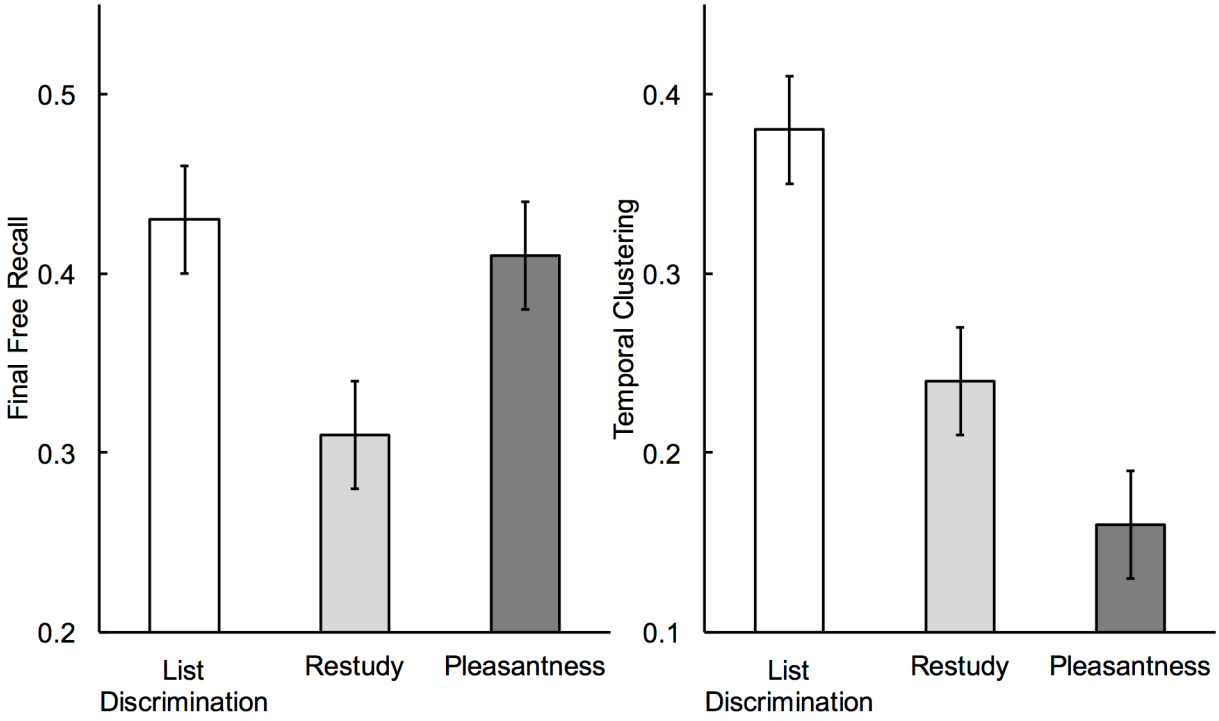

Figure 2. Proportion correct on final free recall and temporal clustering scores in Experiment 2. Error bars represent standard errors of the mean.

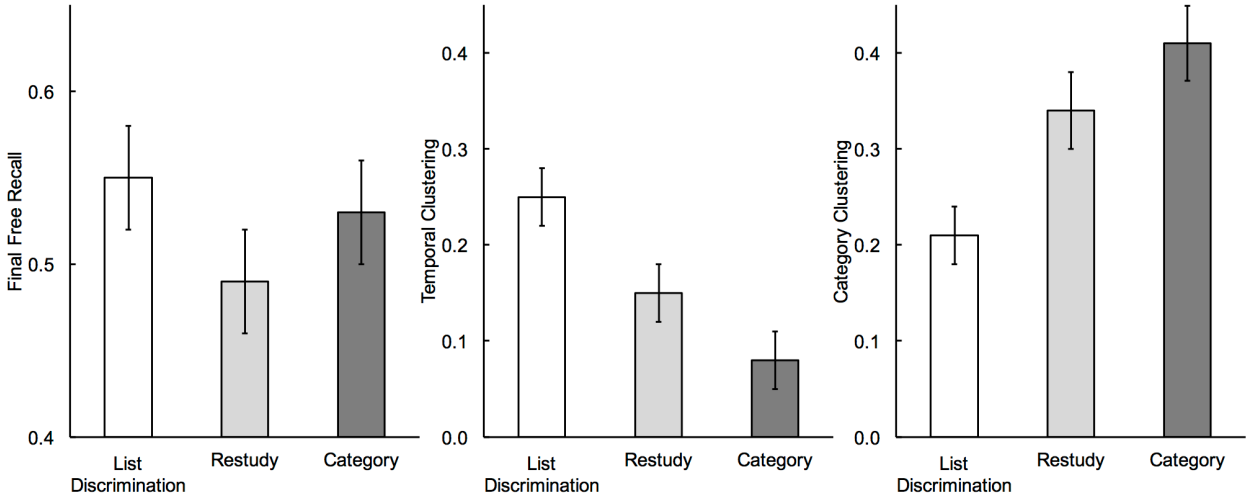

Figure 3. Proportion correct on final free recall, temporal clustering scores, and semantic (category) clustering scores in Experiment 3. Error bars represent standard errors of the mean.