Journal of Experimental Psychology:
Learning, Memory, and Cognition
2008, Vol. 34, No. 1, 112–127

0278-7393/08/$12.00 DOI: 10.1037/0278-7393.34.1.112

Context Retrieval and Context Change in Free Recall: Recalling From Long-Term Memory Drives List Isolation

Yoonhee Jang and David E. Huber
University of California, San Diego

Three experiments used the "list-before-the-last" free recall paradigm (Shiffrin, 1970) to investigate retrieval for context and the manner in which context changes. This paradigm manipulates target and intervening list lengths to measure the interference from each list, providing a measure of list isolation. Correct target list recall was only affected by the target list length when participants engaged in recall between the lists, whereas there were effects of both list lengths with other activities. This suggests that the act of recalling drives context change, thus isolating the target list from interference. Correspondingly, incorrect recall of intervening list items was affected only by the length of the intervening list when recall occurred between the lists, but was otherwise affected by both list lengths. Concurrent with these changes in context similarity, there were apparent changes in context retrieval, as indicated by the overall levels of target retrieval versus intervening recall. A multinomial model of sampling and recovery was implemented to assess the adequacy of this account and to quantify context similarity and context retrieval.

Keywords: free recall, context change, context reinstatement, list-length effect, retroactive interference

Unlike recognition, which asks for a judgment in relation to a specific known item (e.g., do you recognize which car is yours?), recall requires the bringing to mind of information that is not currently at hand (e.g., I remember parking my car at the top level of the garage). Among the vast number of episodes that could be recalled, it is commonly believed that a task-relevant context is used to guide recall, highlighting a subset of possible target memories (e.g., Anderson & Bower, 1972; Estes, 1955; Hintzman, 1988; Howard & Kahana, 2002; Mensink & Raaijmakers, 1988; Murdock, 1997; Raaijmakers & Shiffrin, 1981). In a typical free-recall experiment, participants are presented with a list of items and are asked to recall in any order that they wish from the list just studied. However, this classic paradigm does not necessitate a particular episodic context beyond things that happened recently; this recency context is likely available with little or no effort. In other words, the classic paradigm does not require retrieval of the context itself. By changing the environmental context at the time of study and test, it is known that context similarity is very important in terms of both proactive interference (e.g., Dallett & Wilcox, 1968) and the match of context between study and test (e.g., Godden & Baddeley, 1975; see Smith, 1988, for a review). However, these experiments directly provided an exogenous context for study or retrieval through environmental manipulation and, therefore, have little to say regarding the manner in which endogenous context is generated, changed, and retrieved. In the reported studies, we seek to investigate factors that affect the ability to generate a unique endogenous context and factors that affect the ability to retrieve prior contexts to the exclusion of intervening memories. This is achieved with a retroactive interference paradigm that asks participants to recall not from the most recent list but rather from the list before the last.

It is well known that the proportion of words correctly recalled from a list decreases as the list length increases, which is referred to as the *list-length effect* (e.g., Murdock, 1962; Postman & Phillips, 1965; Roberts, 1972). It is also well known that recall performance for memories previously stored is negatively affected by later learning, which is referred to as *retroactive interference* (e.g., McGeoch & McDonald, 1931; Osgood, 1949). Both of these effects can be viewed as occurring through the use of context. The context used to probe memory highlights items associated with that context, and retrieval takes place within this limited subset of memories. In the case of list length, the other items on the list also match the list context that is used to probe memory, and this produces additional interference in accord with the number of items. In the case of retroactive interference, activities between the to-be-remembered items and final recall may likewise match the probe context and provide additional interference.

This traditional view of context use and interference fails to consider the role of context retrieval itself. Even when there is a highly specific context that is only associated with a subset of target memories (i.e., an isolated list), retrieval or reinstatement of that context may fail, and performance may suffer because of use of an inappropriate context. When testing occurs shortly after a study list is presented, it may not be necessary to reinstate context at all because the study list context matches the current context. However, when there is a sufficient delay filled with intervening events between study and test, reinstatement of the context may be necessary. In the present work, we investigate context reinstate-

This research was supported by National Institute of Mental Health Grant MH063993-04. We thank Sverker Sikström and Petter Kallioinen for their valuable comments and Eddy Davelaar for his helpful early suggestions concerning the design of Experiment 3.

Correspondence concerning this article should be addressed to David E. Huber or Yoonhee Jang, Department of Psychology, University of California, San Diego, 9500 Gillman Drive, La Jolla, CA 92093-0109. E-mail: dhuber@ucsd.edu or yhjang@ucsd.edu.

ment, and in particular we attempt to determine the factors that drive endogenous context changes. Context change may serve to isolate the target list, thereby minimizing retroactive interference, but it may additionally lower performance because target list recall can only be achieved through reinstatement of the target list context. In contrast, if context does not change between lists, then the current context may be more useful for recalling from the target list, even though it produces more interference. By considering both correct recall and incorrect recall for intervening items (i.e., commission errors), we seek to measure these two effects of context change. These may appear to be two mechanisms that produce the same behavioral result and, thus, a distinction without a difference. However, the "list-before-the-last" paradigm (Shiffrin, 1970) can distinguish between these by separately considering list-length effects, which relate to context similarity, and the overall levels of recall for the target list versus intervening recall, which relate to context retrieval. Thus, the notion of "list isolation" is multifaceted and the context associated with an isolated list may uniquely indicate the desired episodic memories, yet, at the same time, isolated memories may be more difficult to recall because they necessarily rely on context retrieval. In terms of a rational analysis of the retrieval process (e.g., Anderson, 1990), it follows that participants will engage in greater context reinstatement for isolated lists, whereas performance may be optimized by using the current context for nonisolated lists.

Shiffrin (1970) developed the list-before-the-last paradigm to determine whether forgetting is due to decay over time (or over experience) or to interference. The experiments of the paradigm involved 20 lists of words presented one after the other, with recall testing occurring between every list. Rather than asking participants to recall from the most recent list, they were told to recall from the list before the last; therefore, the most recent list was the *intervening list* and the list before the last was the *target list*. Because the experiment included a long series of study lists with testing between each list, it was always necessary for participants to study intervening lists considering that later on these lists became the list before the last. To assess interference effects, the number of unique words on both the target list and the intervening list was separately manipulated. The key finding was worse performance for longer target lists, but at the same time the length of the intervening list did not matter (decay theory predicted an effect of the intervening list length). From the standpoint of interference theory, this result indicated that the probe context was selective and specific to the target list (i.e., the target list was isolated); people were perfectly able to focus on the target list to the exclusion of the intervening list.

Global memory models of free recall assume that a probe context is used to focus the retrieval process on the studied list (e.g., Hintzman, 1988; Murdock, 1982; Raaijmakers & Shiffrin, 1981), and for the most part it is assumed that this process is accurate and applied similarly across conditions of interest. Although these models and others presume that there is a single unchanging list context (see also DeLosh & McDaniel, 1996; Rohrer, 1996; Rohrer & Wixted, 1994), other theories allow that the endogenously generated context fluctuates over time or with subsequently presented items, which results in forgetting because of a mismatch between study context and probe context (e.g., Estes, 1955; Howard & Kahana, 1999, 2002; Mensink & Raaijmakers, 1988). For instance, the context fluctuation model of Mensink and Raaijmakers (1988) was designed to augment the search of associative memory (Raaijmakers & Shiffrin, 1981) model of free recall by assuming that context is a vector of elements that is constantly changing as old elements are probabilistically replaced with new elements, thereby causing context to drift, similar to a random walk process. At study, associations between active contextual elements and items are strengthened, and at test, memory is probed using the contextual elements active at that time. Assuming that the probability of retrieval is proportional to the number of contextual elements that are active at the time of both study and test (i.e., the similarity of the two contexts), this model predicts and explains a large number of classic retroactive and proactive interference phenomena and has also been successfully applied to spacing and repetition effects (Raaijmakers, 2003; see also Sirotin, Kimball, & Kahana, 2005).

The temporal context model (TCM), developed by Howard and Kahana (2002), also assumes a continuously changing list context but further specifies the nature of this random fluctuation by assuming that it is at least partially driven by the items themselves. According to TCM, an item retrieves its preexperimental context when it is encountered during study (e.g., under what circumstances is that particular word commonly encountered). This preexperimental context is used to update the previous state of the temporal context (i.e., the context of the list). Through this updating, the temporal context gradually changes during the study list, with the preexperimental context of items early in the list carrying forward to combine with the preexperimental context of items later in the list. For example, one might have an episodic memory of a particular drive to work that involved both eating an apple for breakfast and witnessing a car accident. The memory for this episode is distinct and unique to the individual, and viewing the word *apple* might spark retrieval for episodically related notions of *car* or *accident*. If the subsequent word on the list was *window*, it might be thought of in the context of a car accident, and, thus, an image of a broken window may come to mind. In this manner, an endogenous list context is generated even if the items on the list are not designed to promote a specific exogenous context. Something similar is assumed to take place at retrieval, with the preexperimental context of a successfully retrieved item used to probe subsequent retrieval attempts. This dynamic updating of context is remarkably successful at explaining a complex pattern of data involving recency effects and analyses that are conditional on the position of a retrieved item to examine the probability that neighboring items from the study list are subsequently recalled (Howard & Kahana, 2002; Kahana & Howard, 2005). One of our primary questions in the current research was motivated by the proposal in TCM that the act of retrieving from long-term memory (i.e., preexperimental context) drives context change. In our studies, we assessed this claim directly and ascertained what types of retrieval (e.g., free recall, recognition, semantic recall, and recall from short-term memory) promote context change.

More than 30 years after the original Shiffrin (1970) study, Ward and Tan (2004) used the list-before-the-last paradigm, but with quite different results in some situations. They wondered whether the lack of intervening list-length effects in Shiffrin's study was due to specific strategies, such as rehearsal of the prior list during the intervening list. In their first two experiments, Ward and Tan replicated Shiffrin's finding that recall performance was affected only by the length of the target list and not by the length

of the intervening list. However, in a third experiment, they found effects of both the target and the intervening list lengths. For their first two experiments, Ward and Tan used the original list-before-the-last paradigm (a long series of study lists and tests); this technique mandates a set amount of time devoted to retrieval between each list. However, their third experiment used sets of two lists, with a cue at the end of the second list indicating which list to recall from, rather than presenting an ongoing series of lists with testing between every list. Between the two lists in their third experiment, there was a short 3-s pause to indicate the start of the second list. Following up on this difference between the paradigms, our first experiment sought to replicate both findings within the same experiment by using the original list-before-the-last paradigm with a long series of lists, but randomly intermixing during the sequence of lists a 50/50 mix of intervening recall or no recall between the lists. This within-subjects design ruled out any concern for differential encoding of lists because at the time of study it was unknown whether any particular list was to be followed by recall from the list before the last or whether the list was to be followed by the next list without any need for recalling between the lists.

We report the results from three experiments that systematically manipulated the type of activity between the lists in the list-before-the-last paradigm to determine which activities cause context change (and thus list isolation). Experiment 1 followed up on Ward and Tan's (2004) findings by comparing recall between the lists versus a brief pause. Experiment 2 tested differences of recall testing versus recognition testing between the lists. Finally, Experiment 3 contrasted episodic long-term recall, semantic long-term recall, and retrieval from short-term memory as tasks between the lists. In all three experiments, we also examined incorrect retrieval from the intervening list as an additional measure of list isolation, which was not reported in previous studies. The combination of target recall and intervening recall as a function of both list lengths provides the necessary constraint to separate the effect of context retrieval (which is related to the ratio of correct target recall to incorrect intervening recall) from the effect of context similarity (which is related to the list-length effect of one list vs. the list-length effect of the other list). To quantify both context similarity and context retrieval, we report results from a simple multinomial model as applied to the data of all three experiments. To preview the results, application of the model revealed that recall from long-term memory between the lists served to simultaneously drive context change (thus making the target list context dissimilar) but also makes performance more reliant on context retrieval. Because there are both costs and benefits to list isolation—isolated lists are protected from interference, but performance may suffer depending on context reinstatement success—performance is optimized by different degrees of context reinstatement depending on the level of isolation.

Experiment 1

Experiment 1 was designed to replicate both Shiffrin's (1970) and Ward and Tan's (2004) results within the same experiment by manipulating the type of task between the lists (recall vs. a brief pause) as a within-subjects variable and crossing this manipulation with the length of the target list and the length of the intervening list. Participants in all experiments were presented with a long series of lists, with the number of words per list varying in a pseudorandom fashion. The primary condition of interest was recall of the list prior to the most recent list. Because retrieval was always for the prior list, no list was ever tested more than once. The use of different encoding or retrieval strategies across different conditions should be minimal because all experiments used repeated measures designs that involved a long series of lists, with the ordering of conditions randomly intermixed.

Method

Participants. One hundred twenty undergraduate students at the University of Maryland were recruited and received credit for psychology courses in return for their participation.

Materials. Stimuli consisted of 432 moderately high-frequency (an average of 80 times per million; norms from Kucera & Francis, 1967), singular noun words from three to eight letters in length.

Design. The experimental design was a $2 \times 2 \times 2$ factorial with target list length (short vs. long), intervening list length (short vs. long), and type of task between the lists (recall vs. no test) manipulated within subjects. Recall responses were recoded to provide two dependent measures: the proportion of the target list that was correctly recalled and the proportion of the intervening list that was incorrectly recalled. With this design, there were four list-length combinations: (a) long target list and long intervening list (LL); (b) long target list and short intervening list (LS); (c) short target list and long intervening list (SL); and (d) short target list and short intervening list (SS). Each of the conditions occurred twice with testing between the lists and twice without testing between the lists.

Procedure. Participants were told that they would be presented with a series of short and long lists and that at the end of each list their memory for the list before the last might be tested. They were instructed to recall as many prior list words as they could in any desired order and were additionally told not to recall from the intervening list.

Short lists consisted of 6 words and long lists consisted of 24 words. Participants were first given a practice session with 6 lists, of which 4 were short and 2 were long. Excluding the period following the first list (which could not possibly include testing), testing occurred for 4 of the remaining 5 after-list periods during practice. Following practice, participants were informed that the experimental session was to begin. There were 24 lists during the experimental session, of which 12 were short and 12 were long, with recall testing occurring at the end of 16 of the 24 lists. Because they were not informed which list was the final list, even the final list was presumably studied in preparation for a future test. The assignment of words to lists was randomized anew for each participant. Furthermore, each participant received a different random order of the length and testing conditions. These orders were determined by randomly searching through the possible orders for ones that satisfied the constraint that each of the eight conditions (i.e., four list-length combinations with testing and without testing) occurred exactly twice. As with the practice phase, there was no testing following the first list of the experimental session (because of no target list).

During study, a series of words were presented one at a time for 2.5 s per word, and there was a 2.5-s break following every list.

During this break, participants were either told to prepare for recall testing or that no recall was required and that the next list would begin. When there was a test between the lists, participants were given 45 s to recall from the list before the last. These durations were chosen so that (a) the study–test delay of the LS condition with testing between was equivalent to that of the LL condition without testing between and (b) the study–test delay of the SS condition with testing between was equivalent to that of the SL condition without testing between. We note that Shiffrin's (1970) Experiment 1 used a fixed duration of 60 s for recall, whereas Shiffrin's Experiments 2 and 3 and Ward and Tan's (2004) Experiments 1 and 2 used self-paced recall testing.

Results

Throughout, all tests of statistically significant differences use an alpha of .05, and estimates of effect size are reported as partial eta-squares for significant effects.

Proportion recalled from the correct target list versus proportion recalled from the incorrect intervening list (type of response) was included as a fixed factor in an analysis of variance (ANOVA) of the results so that it could be determined whether the pattern varied as a function of which list a recalled word belonged to. This four-way ANOVA revealed a three-way interaction: Type of Response × Type of Task × Length of Target List, $F(1, 119) = 8.63$, $MSE = .01$, $\eta_p^2 = .07$, and a three-way interaction: Type of Response × Type of Task × Length of Intervening List, $F(1, 119) = 33.57$, $MSE = .01$, $\eta_p^2 = .22$. On the basis of these interactions, we next separately consider the results with recall between the lists versus no test between the lists.

Recall between the lists. Figure 1 (top panel) shows the mean proportion of correct recall and mean proportion of incorrect intervening recall as a function of target and intervening list lengths when a recall test was given between the lists. Correct recall was greater for short target lists than for long target lists, $F(1, 119) = 63.03$, $MSE = .02$, $\eta_p^2 = .35$. There was neither an intervening list-length effect, $F(1, 119) = 1.79$, $MSE = .02$, $p = .18$, nor an interaction for correct recall, $F(1, 119) = 1.06$, $MSE = .01$, $p = .30$. These findings successfully replicate the list-before-the-last results as reported by Shiffrin (1970).

For incorrect intervening recall, there were both target and intervening list-length effects; incorrect intervening recall was greater for short intervening lists than for long intervening lists, $F(1, 119) = 15.64$, $MSE = .003$, $\eta_p^2 = .12$, and was greater for short target lists than for long target lists, $F(1, 119) = 5.57$, $MSE = .002$, $\eta_p^2 = .04$. There was no interaction, $F(1, 119) < 1$.

No test between the lists. Figure 1 (bottom panel) shows the mean proportion of correct recall and mean proportion of incorrect intervening recall as a function of target and intervening list lengths when no test was given between the lists. Both target and intervening list lengths affected correct recall performance; correct recall was greater for short target lists than for long target lists, $F(1, 119) = 164.97$, $MSE = .02$, $\eta_p^2 = .58$, and greater for short intervening lists than for long intervening lists, $F(1, 119) = 46.27$, $MSE = .02$, $\eta_p^2 = .28$, and these two factors interacted, $F(1, 119) = 4.52$, $MSE = .02$, $\eta_p^2 = .04$. These findings are identical to those of Ward and Tan's (2004) Experiment 3 in which they used two lists and a short break between the lists.

Incorrect intervening recall was greater for short intervening lists than for long intervening lists, $F(1, 119) = 8.77$, $MSE = .002$, $\eta_p^2 = .07$, and greater for short target lists than for long target lists,

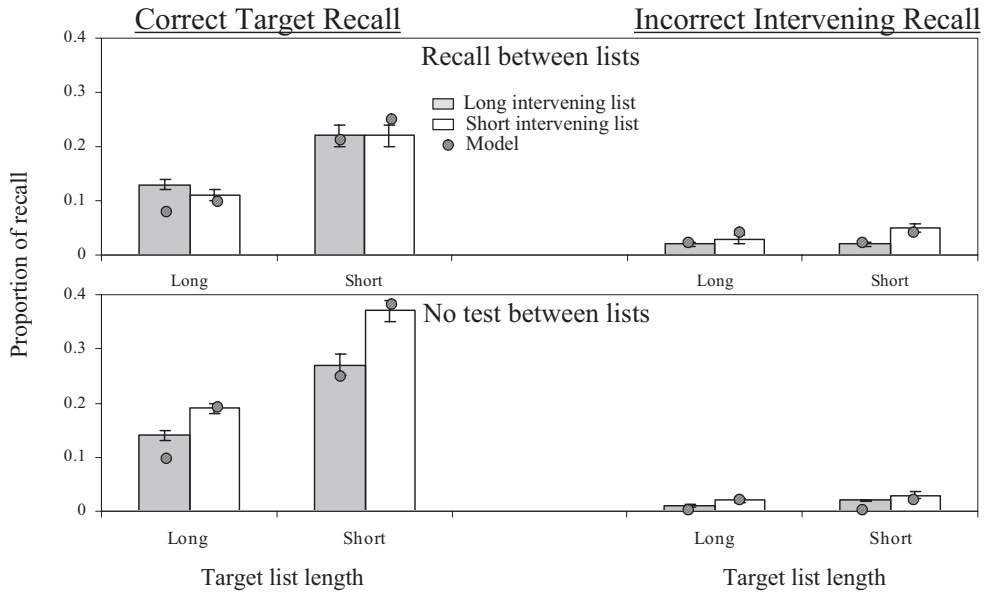


Figure 1. Experiment 1 proportion of list recalled, broken down by correct target recall (left) and incorrect intervening recall (right) as a function of the type of task between the lists (recall vs. no test), target list length (short vs. long), and intervening list length (short vs. long). Error bars depict ±1 standard error of the mean. Dots indicate the multinomial model with the reported best-fitting parameters.

$F(1, 119) = 5.14$, $MSE = .001$, $\eta_p^2 = .04$. There was no interaction, $F(1, 119) = 1.60$, $MSE = .002$, $p = .21$.

Discussion

The results of Experiment 1 replicated those of previous studies, finding no intervening length effect when a recall test was inserted between the lists (Shiffrin, 1970), but a sizable intervening length effect when there was no test between the lists (Ward & Tan, 2004, Experiment 3). It is important that these replications were achieved with a within-subjects design, demonstrating that the key variable is the task that occurred between the lists, rather than other aspects and demand characteristics that may differ between the prior studies. This pattern of results is sensible if one assumes that engaging in recall between the lists served to change the temporal context at the time of memory storage such that the context for the second list was sufficiently different from the first, thus isolating or separating the lists; if a context is used to probe memory that is specific to the target list, then there is minimal interference from the intervening list. In the condition with no test between the lists, the contexts appropriate to each list may have been sufficiently similar so as to produce interference across both lists (i.e., as if it was just one big list). This pattern of context change with testing between the lists is expected both by Howard and Kahana's (2002) TCM, as a result of retrievals driving the change in context, and by Mensink and Raaijmakers's (1988) context fluctuation model, as a result of the passage of time with testing between the lists.

An account of these data based on context change and context retrieval, both of which appear to vary as a function of testing between the lists, provides an adequate qualitative explanation of correct target recall. Context change as a function of the between-list task is highlighted by examining the list-length effects for correct target list recall, noting that there was more intervening list interference for the case of no testing between the lists. Evidence of context retrieval is more subtle and is highlighted in the data pattern with testing between the lists (the upper graph in Figure 1) in which case target recall was affected only by the target list length at the same time that incorrect intervening recall was primarily affected by the intervening list length. This pattern is incommensurate with use of a single specific probe context on all trials. If interference is the direct consequence of the number of items that match the probe context, and the same probe context is used on all trials, then it follows that the list-length effects should be the same for both correct target recall and incorrect intervening recall (i.e., with a single sampling space defined by the probe context, if there is a large target list-length effect and absent intervening list-length effect for correct target recall, then the same should be true for incorrect intervening recall). Instead, the list-length effects were quite different for items recalled from the target list (correct recall) as compared with items recalled from the intervening list (incorrect recall). This pattern could be explained if not every recall attempt used the same context to probe memory; instead, the role of context retrieval may be important, and it may be that some retrieval attempts (or perhaps some testing sessions) failed to reinstate a context appropriate to the target list. If the target context was retrieved, then there was no interference from the intervening list (and no recall from the intervening list). If the target context was not retrieved, then people may have used the current context by default, which would tend to match the intervening list. For this failure to reinstate the target list context, recalled items would come from the intervening list, with minimal interference from the target list.

List isolation appears to produce a trade-off between benefits because of reduced interference and costs because of a greater reliance on context retrieval. This trade-off may explain the specific time-controlled tests built into the design of Experiment 1. Correct recall was greater for the LL condition with no testing between the lists than for the LS condition with testing between, $t(119) = 4.09$, even though these conditions involved the same overall delay between the target list and the time of test. Likewise, correct recall was greater for the SL condition with no testing between than for the SS condition with testing between, $t(119) = 2.26$. To sum up, when comparing conditions that control the lag between the end of the target list and the start of recall testing, accuracy was greater when the intervening period was filled with more study items than with fewer study items and an intervening testing session. These results may seem paradoxical considering that the pattern of results indicates that testing between the lists helped to differentiate the lists. However, it is important to realize that there are two routes for recall of an item from the target list; either the target list context is properly reinstated, thus filtering out the intervening list, or context is not reinstated but the current context is sufficiently similar to allow retrieval of target list items (as well as retrieval of intervening list items). If context reinstatement is difficult such that this second retrieval route dominates, then performance is actually helped by similarity between the list contexts, even though this produces more interference from the intervening list.

Finally, it appears that the probability of target list context retrieval varied as a function of task between the lists because there was not much of a change in overall performance levels (i.e., collapsing across list lengths) in comparing recall testing versus no testing between the lists, even though recall testing between the lists was presumably more reliant on context retrieval. Because intervening list recall was unaffected by the target list length in the case of recall testing between the lists, this suggests that use of the current context at the time of testing was completely ineffectual for target list retrieval (i.e., it only matches the intervening list). Thus, one might expect target performance to be substantially lower with recall testing between the lists because performance fully relied on successful context retrieval. Yet, as seen in Figure 1, performance was approximately the same in comparing recall testing versus no testing between the lists, suggesting an adaptive modification in the context used to probe memory. When the target list context was unique, retrieval more heavily used that context, but when the target list context still sufficiently matched the current context, retrieval more heavily used the current context.

Experiment 2

In Experiment 1, inserting recall testing between the lists apparently promoted context differentiation and protected each list from interference from the other list. Experiment 2 was designed to replicate these findings with the shorter 1-s study item duration originally used by Shiffrin (1970) and additionally ascertain whether context differentiation occurs only with recall testing or whether recognition testing between the lists likewise serves to promote context change.

Method

Participants. One hundred undergraduate students at the University of Maryland were recruited and received credit for psychology courses in return for their participation.

Materials. The 432 words from Experiment 1 were used and augmented with an additional 158 words using the same selection criterion as Experiment 1.

Procedure. The procedure was identical to that of Experiment 1 except as noted. Participants were told that their memory of the list before the last would be tested either with forced-choice recognition or with free recall and that there was no way to know in advance which type of test they would receive in relation to any particular list. Each of the four conditions (i.e., LL, LS, SL, and SS) was presented eight times per participant across 32 study lists, allowing four recall tests and four recognition tests for each condition, each of which further broke down as containing two instances with recognition testing between the lists and two instances with recall testing between the lists. In total, there were 34 study lists, although testing of the first study list was not included in the analyses (there was no testing between the first 2 lists), and the last study list was not included in the analyses (it was never tested). Study words appeared at a rate of 1 s per item. During recall testing, participants were given 60 s to type their recall responses. Recognition testing was self-paced and included four forced-choice trials. Of these four trials, two used foils selected from the intervening list, and the other two used foils that were new within the experiment. For recognition, all six words from a short study list were used (four as targets and two as intervening list foils), and the first six words from a long study list were likewise used. In light of this design with a maximum of four recognition trials, it was not possible to mandate that recognition testing take as long as recall testing (doing so would leave plenty of idle time and possibly promote rehearsal strategies). On average, participants took 8 s to engage in recognition testing. In any event, Experiment 3 provides further replication of the difference between different types of testing in a time-controlled manner.

Results

Experiment 2 was originally designed as a direct comparison between recognition and recall performance in the list-before-the-last paradigm rather than a study of recall as a function of recognition versus recall testing between the lists. For simplicity and consistency across the experiments, we consider only the recall results, although Appendix A reports the recognition results and statistical tests on those data.

A four-way ANOVA on the recall data revealed a four-way interaction, $F(1, 99) = 9.39$, $MSE = .01$, $\eta_p^2 = .09$; a three-way interaction: Type of Response $\times$ Type of Task $\times$ Length of Intervening List, $F(1, 99) = 55.33$, $MSE = .01$, $\eta_p^2 = .34$; and another three-way interaction: Type of Task $\times$ Length Target List $\times$ Length of Intervening List, $F(1, 99) = 30.39$, $MSE = .01$, $\eta_p^2 = .24$. Next, we report the results separately for recognition and recall between the lists.

Recall between the lists. Figure 2 (top panel) shows the mean proportion correct recall and mean proportion incorrect intervening recall as a function of target and intervening list lengths when recall was given between the lists. Correct recall was greater for short target lists than for long target lists, $F(1, 99) = 14.26$, $MSE = .01$, $\eta_p^2 = .13$. There was no effect of intervening list length, $F(1, 99) = 3.37$, $MSE = .01$, $p = .07$, and no interaction between intervening list length and target list length, $F(1, 99) =$

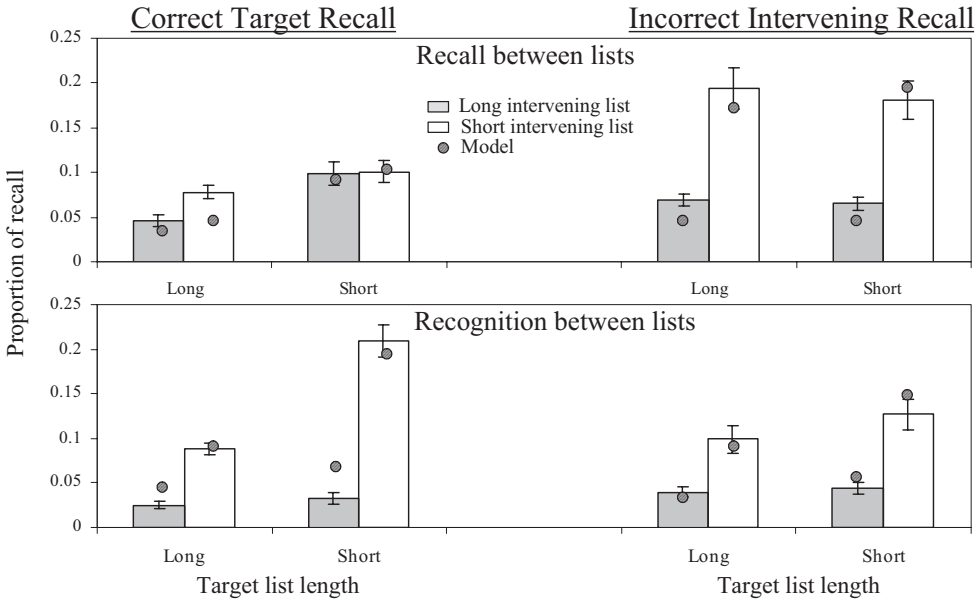


Figure 2. Experiment 2 proportion of list recalled, broken down by correct target recall (left) and incorrect intervening recall (right) as a function of the type of task between the lists (recall vs. recognition), target list length (short vs. long), and intervening list length (short vs. long). Error bars depict ± 1 standard error of the mean. Dots indicate the multinomial model with the reported best-fitting parameters.

2.59, $MSE = .01$, $p = .11$. These findings replicate Shiffrin (1970) and our Experiment 1 results.

Incorrect intervening recall was greater for short intervening lists than for long intervening lists, $F(1, 99) = 59.36$, $MSE = .02$, $\eta_p^2 = .38$, and there was no effect of target list-length effect and no interaction between target list length and intervening list length, $F(1, 99) < 1$ for both.

Recognition between the lists. Figure 2 (bottom panel) shows the mean proportion of correct recall and mean proportion of incorrect intervening recall as a function of target and intervening list lengths when recognition was given between the lists. Correct recall was greater for short target lists than for long target lists, $F(1, 99) = 45.97$, $MSE = .01$, $\eta_p^2 = .32$, and was also greater for short intervening lists than for long intervening lists, $F(1, 99) = 128.64$, $MSE = .01$, $\eta_p^2 = .56$. Furthermore, the two list lengths interacted, $F(1, 99) = 29.75$, $MSE = .01$, $\eta_p^2 = .23$. These findings are identical to those of Ward and Tan's Experiment 3 (2004) and of our Experiment 1 in which there was no test between the lists.

Incorrect intervening recall was greater for short intervening lists than for long intervening lists, $F(1, 99) = 35.22$, $MSE = .01$, $\eta_p^2 = .26$. There was neither a target list-length effect, $F(1, 99) = 3.38$, $MSE = .01$, $p = .07$, nor an interaction between target list length and intervening list length, $F(1, 99) = 2.32$, $MSE = .01$, $p = .13$.

Discussion

Experiment 2 produced the same qualitative data patterns as Experiment 1, even though the study time was substantially shorter (1 s), which produced overall lower correct recall from the target list and greater incorrect recall from the intervening list (as seen in Figure 2, there was just as much intervening recall as correct recall). Furthermore, the change in the pattern from one of list isolation to one of blending across the lists occurred as a function of recall versus recognition testing between the lists rather than recall versus no testing. This suggests that (a) episodic recall more effectively causes change in the temporal context (although note that recognition testing took only 8 s on average) and (b) in general these patterns hold across situations involving both strong (Experiment 1) and weak (Experiment 2) memories, as might be expected if it is the context at study that is the key underlying variable rather than memory strength.

Experiment 2 found additional evidence of context similarity and context retrieval effects, as well as variations in these effects as a function of between-list task. With recall between the lists, there was no effect of manipulating the intervening list length on target list recall, and, conversely, there was no effect of manipulating the target list length on intervening list recall. In other words, a recalled item was affected only by the length of the list to which it belonged. This is sensible if some retrieval attempts used a context that was unique to the target list and others used a context that was unique to the intervening list (i.e., a recency context). Again, this demonstrates the need for context retrieval. Conversely, with recognition between the lists there were equal effects of both list lengths for both target list recall and intervening list recall, which suggests that the endogenous contexts associated with each list were very similar. Despite the nearly complete list isolation with recall testing between the lists, which suggests a greater reliance on context retrieval to recall from the target list, performance was roughly the same regardless of task between the lists. In other words, it appears that context retrieval was used more extensively for isolated lists.

Experiment 3

Experiments 1 and 2 found evidence that endogenous context changes with recall but not with recognition or no testing. However, it is not clear whether context change occurs only with episodic recall or whether it occurs more generally with other kinds of recall, such as recalling generic knowledge facts or recalling from short-term memory. Moreover, free recall from the list before the last is a very difficult task, and it may be that effort (i.e., task difficulty) is the key underlying variable behind context change. Finally, one lingering concern for the results of Experiment 2 is that the comparison between recognition and recall was not equated in task duration. Experiment 3 addresses all these questions by comparing three different tasks between the lists: (a) 60 s of difficult episodic free recall from the list before the last (the standard condition); (b) 60 s of easy recall from lexical-semantic long-term memory (letter completion task); and (c) 60 s of difficult recall from short-term memory (2-back task). If any kind of recall drives context change, or if only time between the lists matters, then all three tasks will reveal list isolation. If it is task difficulty that matters, then only Tasks 1 and 3 will reveal list isolation. If it is recall from long-term memory (either episodic or semantic) that matters, then only Tasks 1 and 2 will reveal list isolation. Finally, if it is only episodic recall that matters, then only Task 1 will reveal list isolation.

Method

Participants. Fifty-nine undergraduate students at the University of Maryland and 96 undergraduate students at the University of California, San Diego (i.e., 155 participants) were recruited and received credit for psychology courses in return for their participation.

Materials. The words from Experiment 2 were used to fill the study manipulations. For the letter completion task, 50 words from 5 to 10 letters in length that contained an *i* (e.g., *alligator*) and 50 words that contained an *e* (e.g., *algebra*) were used, with the *i* or *e* removed (e.g., *all_gator* or *alg_bra*). These words did not appear as study words. For 2-back task, letters were used.

Procedure. The procedure was identical to that of Experiment 1 except as noted. During the letter completion task, participants were shown a series of words, one at a time for 3 s each, indicating whether the missing letter was an *i* or *e*. Twenty words were tested across the 60 s, with 10 missing *i* and 10 missing *e*. During the 2-back task, participants were presented with a sequence of letters, one at a time, and were required to press a key to indicate whether the current letter was the same as the letter before the last (i.e., the letter presented two positions back in the sequence). Forty letters were shown, with each letter presented for 1.5 s. On average, 12.5% of the trials presented letters that were the same as 2-back. Feedback was provided on every trial of both tasks.

During the practice phase, there were 2 recall tests, 2 letter completion tests, and 2 2-back tests (6 lists total). During the experimental phase, there were 12 recall tests, 5 letter completion

tests, and 5 2-back tests (22 lists total). After presentation of List 1 in both phases, there was either a letter completion task or a 2-back task. The remaining lists were followed by any of the three possible tasks in pseudorandom order. Study duration was set at 1.75 s per item (which was the average duration used across Experiments 1 and 2). Each of the four list-length conditions (LL, LS, SL, and SS) was presented once in combination with each of the three between-list tasks.

Results

Performance of letter completion task (M = .94, SEM = .004) was greater than that of 2-back task (M = .74, SEM = .018), $t(154) = 11.31$, demonstrating that the 2-back task was more difficult.

For list-before-the-last recall, there was a four-way interaction, $F(2, 308) = 12.53$, $MSE = .02$, $\eta_p^2 = .08$. All four three-way interactions were significant: Type of Task × Length of the Target List × Length of the Intervening List, $F(2, 308) = 11.53$, $MSE = .02$, $\eta_p^2 = .07$; Type of Response × Type of Task × Length of the Target List, $F(2, 308) = 12.86$, $MSE = .02$, $\eta_p^2 = .08$; Type of Response × Type of Task × Length of the Intervening List, $F(2, 308) = 4.40$, $MSE = .02$, $\eta_p^2 = .03$; and Type of Response × Length of the Target List × Length of the Intervening List, $F(1, 154) = 11.94$, $MSE = .02$, $\eta_p^2 = .07$. Next, recall is reported separately as a function of the task between the lists. Performance in the letter completion and 2-back tasks is reported in Appendix B.

Recall between the lists. Figure 3 (top panel) shows the mean proportion of correct recall and mean proportion of incorrect intervening recall as a function of target and intervening list lengths when recall was given between the lists. As in Experiments 1 and 2, this condition replicated the findings of Shiffrin (1970); correct recall was greater for short target lists than for long target lists, $F(1, 154) = 5.99$, $MSE = .02$, $\eta_p^2 = .04$, and there was neither an intervening list-length effect, $F(1, 154) = 2.11$, $MSE = .02$, $p = .15$, nor an interaction, $F(1, 154) = 3.74$, $MSE = .02$, $p = .06$.

Analogously, incorrect intervening recall was greater for short intervening lists than for long intervening lists, $F(1, 154) = 17.03$, $MSE = .01$, $\eta_p^2 = .10$. There was neither a target list-length effect, $F(1, 154) < 1$, nor an interaction between the two list lengths, $F(1, 154) < 1$.

Letter completion task between the lists. Figure 3 (middle panel) shows the recall results when a letter completion task was given between the lists. As with recall between the lists, correct

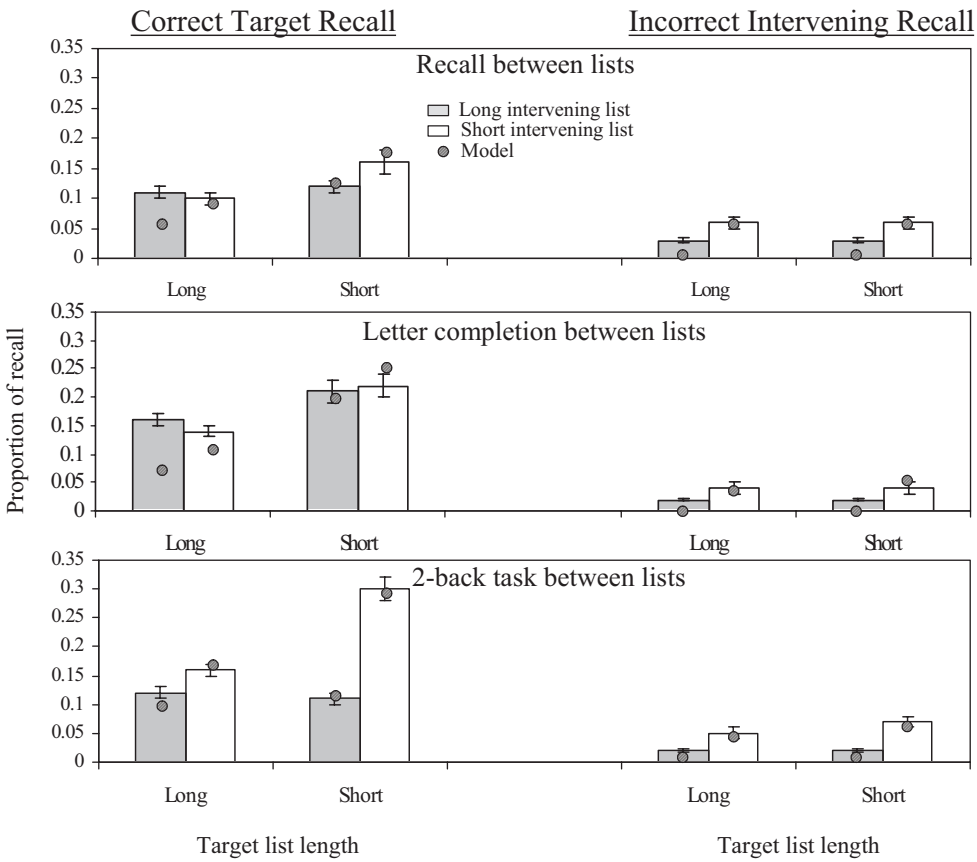


Figure 3. Experiment 3 proportion of list recalled, broken down by correct target recall (left) and incorrect intervening recall (right) as a function of the type of task between the lists (recall, letter completion, or 2-back). Error bars depict ±1 standard error of the mean. Dots indicate the multinomial model with the reported best-fitting parameters.

recall was greater for short target lists than for long target lists, $F(1, 154) = 19.01$, $MSE = .04$, $\eta_p^2 = .11$, but there was neither an intervening list-length effect, $F(1, 154) < 1$, nor an interaction between the list lengths, $F(1, 154) = 1.67$, $MSE = .03$, $p = .20$.

Incorrect intervening recall was greater for short intervening lists than for long intervening lists, $F(1, 154) = 13.61$, $MSE = .01$, $\eta_p^2 = .08$, but there was neither a target list-length effect, $F(1, 154) < 1$, nor an interaction between the list lengths. Because this pattern is identical to that of the results with episodic recall between the lists, it suggests that any kind of recall from long-term memory between the lists is sufficient to produce list isolation.

Two-back task between the lists. Figure 3 (bottom panel) shows the recall results when a 2-back task was given between the lists. Unlike the previous two task conditions, there was both a main effect of target list length, $F(1, 154) = 20.30$, $MSE = .03$, $\eta_p^2 = .12$, and a main effect of intervening list length, $F(1, 154) = 72.63$, $MSE = .03$, $\eta_p^2 = .32$; in both cases, correct recall was greater for shorter lists than for longer lists. These list-length effects interacted, $F(1, 154) = 37.96$, $MSE = .03$, $\eta_p^2 = .20$.

Incorrect intervening recall was greater for short intervening lists than for long intervening lists, $F(1, 154) = 23.51$, $MSE = .01$, $\eta_p^2 = .13$, and there was neither a target list-length effect, $F(1, 154) = 2.36$, $MSE = .01$, $p = .13$, nor an interaction between the list lengths, $F(1, 154) = 1.62$, $MSE = .01$, $p = .21$. Because the 2-back task is difficult, the difference between this pattern and the patterns seen with episodic recall and semantic missing letter recall demonstrates that neither task difficulty nor short-term memory recall are sufficient to drive list isolation.

Discussion

Experiment 3 found that the act of recalling from long-term memory between the lists (i.e., both episodic list-before-the-last recall and semantic missing letter recall) produced patterns of data consistent with list isolation, whereas the difficult 2-back short-term memory produced blending across the lists, similar to the results with no testing or a short period of recognition testing. Such findings are in agreement with TCM (Howard & Kahana, 2002) if one assumes that retrieval from short-term memory in the 2-back task does not involve retrieval of the associated preexperimental context. For instance, it may be that short-term memory does not involve context at all or that the context in short-term memory includes only the current context without updating from preexperimental context. In contrast, storage and retrieval from episodic memory or from lexical-semantic memory may involve access of the preexperimental context that serves to drive context change. This possibility also explains the failure to change context with recognition testing observed in Experiment 2 under the commonly adopted assumption that recognition retrieval primarily involves a nonspecific familiarity signal rather than a retrieval of a specific context (although note that this might differ if participants engage in recollection-based recognition).

Beyond context similarity effects, Experiment 3 replicated the need for context retrieval. For the recall task and letter completion task, there was only a list-length effect in relation to the list of the recalled items, suggesting that sometimes retrieval used a probe context specific to the target list and other times retrieval used a probe context specific to the intervening list. In contrast, for the 2-back task, there were effects of both list lengths for correct recall, although the effect was larger as a function of the intervening list, suggesting that although the probe context matched both lists, it matched the intervening list more so. Finally, note that target performance was not much worse for isolated lists, suggesting a greater use of context retrieval with isolated lists. Next, we implement a simple multinomial model to assess the adequacy of context similarity, context retrieval, and variations in these mechanisms as a function of between-list task.

A Multinomial Model of Recall

The data from all three experiments appear to require (a) some intermediate degree of context similarity between the target and intervening lists to produce an appropriate level of interference from the other list; (b) context retrieval such that sometimes the probe context matched the target list and other times it matched the intervening list; and (c) variations in these settings as a function of list isolation. Simultaneously accounting for correct target list recall and incorrect intervening list recall is highly constraining, and, as seen in Figures 1–3, the 16 (Experiments 1 and 2) or 24 (Experiment 3) conditions all varied greatly. Initial attempts at fitting these highly constraining data patterns revealed the need for an additional mechanism beyond context similarity and context retrieval. The additional mechanism is highlighted by the 2-back task in Experiment 3, shown in the lowest panel of Figure 3 (although something similar is seen for Experiments 1 and 2). In this nonisolated case, target list recall was actually more affected by the length of the intervening list than by the length of the target list. It suggests a high degree of similarity in the context of the two lists and a relative failure to retrieve the target list context. However, one would expect this situation to produce higher levels of intervening recall as compared with target recall. Yet, quite clearly the opposite is true, and people produced relatively few commission errors from the intervening list, even though there was a substantial intervening list effect. It strongly suggests that some sort of censoring or filtering process took place such that retrieved intervening list items were less likely to be overtly produced. The current data do not sufficiently constrain the mechanism behind this censoring, although it may occur, for instance, by filtering out items that still lie within short-term memory (which would indicate that they come from the intervening list) or, perhaps, through a recognition to reject strategy following retrieval (e.g., Anderson & Bower, 1972).

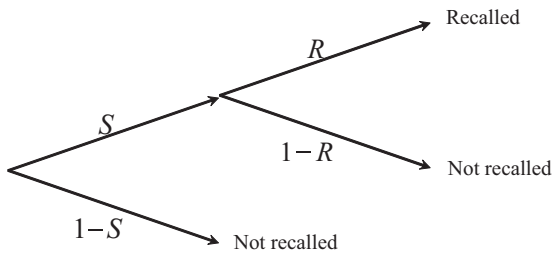
To quantify context retrieval, context similarity, and the censoring of the intervening list, we developed a simple multinomial model with three corresponding decision branches that selected the probe context (target vs. intervening), selected an item that matched the probe context (depending on the similarity between contexts), and overtly recovered or produced that item with some probability depending on whether the item came from the target list or the intervening list. The details of this model are described next.

Interference from other items from the same list and possibly from items from other lists is realized through a sampling space that includes items that match the probe context. Among these context-matching items, particular memories are stochastically chosen assuming uniform sampling (i.e., the probability of sampling an item, S, is equal to the inverse of the number of items contained in the sample space). Even if an item is sampled,

successful recall additionally requires that a sampled item is recovered, which occurs with probability *R*. Figure 4 shows a multinomial model (e.g., Batchelder & Riefer, 1990) of sampling (*S*) and recovery (*R*) that describes one retrieval attempt in relation to a particular item within the sample space. During each testing session, participants make more than one retrieval attempt, which is captured by the parameter *k*, representing the number of times that the sampling and recovery are reiterated with this multinomial model. The equation seen in Figure 4 represents the probability of recalling a particular item contained in the sampling space under the simplifying assumptions of independent decision branches and a sampling process that occurs with replacement following each of the *k* retrieval attempts (i.e., the equation is the probability of recalling any particular item at least once).

We implemented competitive sampling followed by item-specific recovery in this generic manner such that the model is largely consistent with most global memory models (e.g., Dennis & Humphreys, 2001; Gillund & Shiffrin, 1984; Hintzman, 1988; Mensink & Raaijmakers, 1988; Murdock, 1982; Raaijmakers & Shiffrin, 1981; Shiffrin & Steyvers, 1997). All of these models have in common the claim that a probe context is used to define a sampling space from which an item is probabilistically sampled (Luce, 1977), but that sampling itself is not sufficient and must be followed up by successful recovery to fully recall an item. The Luce choice rule states that the probability of sampling is the strength of a memory relative to the strength of all the memories in the sampling space (i.e., it is the process that includes interference from other memories). Once a memory is sampled, it is as if the memory is only partially retrieved (such as in tip of the tongue) and not necessarily ready for the full production of the item's label that is required by recall instructions. Subsequent to sampling, the recovery process depends on the absolute strength of the memory rather than the relative strength.

It is commonly assumed that sampling is limited only to the appropriate list items, but our results require that items from the other list enter the sampling space. We capture this behavior with a context mixing parameter, *M*, representing the probability that an item from the other list enters the sampling space. This can be conceptualized by the context pie charts in Figure 5, representing



$$p(recalled) = 1 - (1 - SR)^k$$

Figure 4. Probability of sampling (*S*) and probability of recovery (*R*) implemented in a multinomial decision tree. For each retrieval attempt, an item is recalled only if it is both sampled and recovered. Assuming sampling with replacement, the equation is the probability of recall over all *k* retrieval attempts. This decision tree is defined in relation to a specific list item. If the same parameters hold true for all other items from the same list, this equation calculates the predicted proportion recalled.

use of a context that is appropriate to all the target list items, with *M* representing the proportion (on average) of the intervening list items that also enter the sampling space defined by the probe context. Conversely, if a context is used that is appropriate to the intervening list, then all the intervening list items enter the sampling space as well as the proportion *M* of the target list items. The larger the mixing quotient (ranging from 0.0 to 1.0), the more the other list is included in the sampling space, and the more recall will reflect not only a length effect from the list that an item comes from, but also a length effect of the other list. In this manner, the model implements context similarity.

Even when there was adequate list isolation and only length effects in relation to the list of the retrieved item, the data revealed substantial incorrect intervening recall (e.g., as seen in Experiment 2). An appropriately low value for *M* can capture the relative list length of effects of each list, but context reinstatement is additionally needed to capture the proportion of trials that result in a context that is more similar to the target list as compared with the proportion of trials that result in a context that is more similar to the intervening list. This is captured through an initial decision branch (see Figure 5), prior to sampling, with this branch representing the probability that the probe context is appropriate to the target list, C_T, or by default use of a probe context that is appropriate to the just-studied intervening list, $C_I = 1 - C_T$. In other words, people attempt target list reinstatement, but if they fail to reinstate, then they use the current context, which will match the intervening list. In this manner, the model implements context retrieval.

Because the degree of mixing seen in the data is often incommensurate with the ratio of correct recall versus incorrect intervening recall, we allow a separate recovery probability for target list items, R_T, versus intervening list items, R_I. To capture the data, it is expected that R_I will be smaller that R_T, but this does not imply that more recent intervening list items are harder to recover; instead, items from both lists are probably equally difficult to recover, but successful recovery of intervening list items might be followed by a subsequent filtering process, perhaps as guided by short-term memory. Filtering of this sort is mathematically identical to two different recovery probabilities, and so we do not bother to explicitly include an additional branch for filtering. This can also be viewed as a "recall to reject" strategy, and it is captured by comparing the recovery rates for the items of each list. In this manner, the model implements item censoring.

To capture these three mechanisms, there are three free parameters, *M*, C_T, and R_I, and to capture the overall level of performance for a particular experiment, R_T is needed. L_T and L_I are constants set to 6 or 24, depending on the simulated list-length condition. As seen in Figure 5, $S_T{}^T$ is the probability of sampling a target item given that the target context is chosen, $S_T{}^I$ is the probability of sampling a target item given that the intervening context is chosen, $S_I{}^T$ is the probability of sampling an intervening item given that the target context is chosen, and $S_I{}^I$ is the probability of sampling an intervening item given that the intervening context is chosen. These intermediate calculations are determined by the list-length constants and the free parameter *M*. Assuming independence in the decision tree branches, the *SR* joint probability of sampling and recovery for correct target recall or incorrect intervening recall is found by multiplying through the two branches that result in each type of recovery, and adding these two

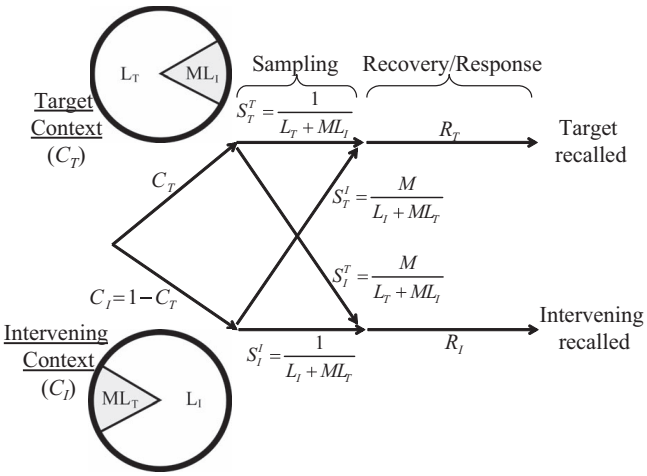


Figure 5. A simple multinomial model of context retrieval, context similarity, and censoring. The decision tree in Figure 4 is expanded to specify the nature of sampling by including a first branch of which context is chosen to probe memory (a context appropriate to the target list or one appropriate to the intervening list) and a second branch of which item is sampled given that context (sampling of a target item vs. sampling of an intervening item). Context similarity is implemented with a mixing parameter, M, representing the probability that an item from the other list enters the sampling space; context retrieval is implemented through the probability of choosing the target list context, C_T; and censoring of intervening items is implemented by comparing the probability of target item recovery to the probability of intervening item recovery (R_I/R_T). Experimental constants: L_T = length of target list; L_I = length of intervening list. Intermediate calculations: $S_T{}^T$ = sampling of an item of target list given target context; $S_I{}^I$ = sampling of an item of intervening list given intervening context; $S_I{}^T$ = sampling of an item of intervening list given target context; $S_T{}^I$ = sampling of an item of target list given intervening context. Assuming independence between branches, the total sampling–recovery probability (*SR*) is derived separately for correct recall and intervening recall by multiplying through the branch probabilities and adding the two branches that result in correct target recall or the two that result in incorrect intervening recall. These *SR* values are then plugged into the equation in Figure 4 to capture the cumulative probability of recall over k attempts.

branch path values. The resultant *SR* joint probability summed across paths is then entered in the equation seen in Figure 4 to determine the probability of recall across all k retrieval attempts.

Fitting the model to the results consisted of the four basic parameters, C_T, M, R_T, and R_I, as well as allowing no additional parameter, just one parameter, or two parameters to vary to capture the effect of the task between the lists (k was set to 15 throughout and not estimated[1]). For example, when the M value varied with task between the lists, each of the first two experiments needed C_T, R_T, R_I, and two Ms, one for recall and the other for no testing (Experiment 1) or recognition (Experiment 2), and Experiment 3 needed C_T, R_T, R_I, and three Ms for recall, letter completion, and the 2-back task. When both M and C_T values varied with task between the lists, each of the first two experiments required six free parameters (two C_T values, two M values, R_T, and R_I), and Experiment 3 required eight (three C_T values, three M values, R_T, and R_I). Considering that at most six free parameters were allowed for experiments with 16 conditions and at most eight free parameters were allowed for the experiment with 24 conditions, these were highly constrained model fits.

The sum of the chi-squares, which is a maximum likelihood estimator, was used to estimate free parameters. Unlike least squares, chi-square places larger importance on small degrees of error that are closer to the accuracy extremes of 0 or 1 and also assigns error in proportion to the number of samples (i.e., longer lists have larger Ns, placing four times as much importance on the results from long lists). In light of the greater importance assigned to misfits of long lists and to misfits of intervening recall, which was closer to 0, we could have fit these data using a least squares technique, which would have produced fits that appeared even closer to the data in the figures (i.e., visual examination of the fits does not adjust for different Ns in different conditions or for the magnitude of error near 0). Nonetheless, the fits appearing in Figures 1–3, which allowed both M and C_T values to vary with between-list task, seem reasonable. Therefore, we felt it was important to use the more statistically justified method of chi-square error fitting so as to produce parameter estimates that resulted from maximum likelihood estimation.

All model fits were significantly bad in terms of chi-square goodness of fit, and all fits that included additional parameters were significantly better, which is hardly surprising given the huge Ns involved in these experiments (e.g., 155 people $\times$ 24 words in a list = 3,720). Although chi-square was used to fit the data, Table 1 presents the average squared error per condition for the raw proportion correct values. This was done to provide a goodness-of-fit measure that was not conflated with the number of observed data points (i.e., a measure that places equal importance on short and long lists and equal importance on the results of each experiment regardless of the number of participants and trials). In this manner, the separate goodness-of-fit values in Table 1 are directly comparable within and across experiments. The four columns from the righthand side of Table 1 ($M + C_T$, M, C_T, and *None*, respectively) show these error values separately when allowing each of the labeled parameter(s) to take on different values to capture the effect of between-list task. The None column did not allow any parameter to vary as a function of between-list task, providing a comparison value to assess the degree of relative improvement for each parameter. For every experiment, allowing M to vary to capture the effect of the task between the lists produced the best fit when allowing just a single parameter to vary (it was also true for chi-square error). In other words, changes in context similarity provided the largest single form of improvement as compared with assuming no differences as a function of between-list task. In addition, as seen in the $M + C_T$ column, allowing both context similarity and context retrieval to simultaneously change as a function of between-list task produced further improvements over each mechanism in isolation, particularly for Experiments 1 and 3.

In light of these results, Table 2 gives parameter values that produced the best model fit with both M and C_T varying as a function of task between the lists. The behavior of the model with these best-fitting parameters is shown in Figures 1–3 (the dots on each bar). Sensibly, the M values seen in Table 2 reveal that no testing (Experiment 1), recognition (Experiment 2), and the 2-back task (Experiment 3) between the lists produced more mixing

[1] We varied the parameter k from 3 to 100, finding no major differences in goodness of fit. Therefore, k was arbitrarily set to 15 for all the reported simulations.

Table 1
Goodness of Fit Shown With Squared Error per Condition Multiplied by 10,000

Experiment	Parameter varied with task between the lists			
	$M + C_T$	M	C_T	None
1	4.13	5.75	6.88	13.88
2	4.06	4.19	8.69	10.00
3	7.79	10.21	10.79	14.63

Note. M = proportion of shared context (mixing); C_T = probability of target context retrieval.

(higher M) in the sample space such that the words from the other list were more likely to interfere with sampling. Higher M values correspond to a greater degree of similarity between the separate contexts associated with each of the lists. In addition, the C_T values seen in Table 2 reveal that episodic recall (all experiments) as well as lexical-semantic recall (Experiment 3) between the lists produced more retrieval of the target context (higher C_T), revealing a greater reliance on context reinstatement for situations that promoted list isolation. In other words, there appears to be a consistent reciprocal relationship between these two parameters both within and across experiments: List isolation produced both greater contextual dissimilarity between the lists (lower M) and greater retrieval and reinstatement of the target list context (higher C_T). In terms of optimizing target list retrieval, this reciprocal relationship is sensible. With dissimilar contexts, the current context provides a poor match to the target list, and so performance is maximized by engaging in the potentially costly act of context reinstatement. In contrast, with similar contexts, the current context works well enough, even though it tends to produce more intervening list intrusions. This supports the idea that retrieval takes place in the most efficient way possible, and a context of convenience (i.e., the current context) is used unless the situation (i.e., list isolation) demands the more effortful act of context reinstatement.

Regarding the absolute magnitude of the parameters, successful retrieval of the target context (i.e., context reinstatement) was in principle quite difficult, as revealed by the low values of C_T. However, as a result of censoring (e.g., R_T values that were much higher than R_I), these low C_T values nevertheless produced more target recall than intervening recall in most instances. In other words, people were often unable to reinstate the appropriate target context, but nevertheless the current context at least partially matched the target list, producing adequate levels of target recall. Because most retrieval used the current context, which matched the intervening list more strongly, censoring of the recalled intervening items was needed to reduce the intervening list intrusion rate. This censoring of the intervening list was particularly pronounced in Experiment 1, which used the longest study duration (2.5 s rather than 1 or 1.75 s for Experiments 2 and 3). With longer study time, the list context may have been more fully and accurately encoded in Experiment 1. Thus, the source of retrieved intervening list items was more readily available for Experiment 1, resulting in greater censoring.

General Discussion

The present study used the list-before-the-last paradigm, replicating the results of both Shiffrin (1970) and Ward and Tan (2004, Experiment 3), thus demonstrating that the difference between these studies was due to the presence or absence of recall testing between the lists. Beyond replicating both sets of results in a within-subjects design, the reported experiments investigated additional between-list tasks that served to reduce interference from an intervening study list. The three reported experiments found that episodic recall or lexical retrieval between the lists served to reduce intervening list interference, whereas a short period of recognition testing, 60 s of a difficult short-term memory task, or no task between the lists resulted in nearly equal interference from both lists. List interference was assessed by manipulating the list length of the target list and the list length of the intervening list. Between-list tasks that promoted list isolation resulted in target list-length effects but no intervening list-length effects, whereas tasks that produced blending across lists resulted in list-length effects from both lists. As predicted by Howard and Kahana's (2002) TCM, these results indicate that recall of information from long-term memory serves to drive context change, thus isolating a prior list. Unlike traditional experiments that test recall immediately or use explicit exogenous manipulations of context, these experiments indicate that retrieval from long-term memory is a key factor in the development of endogenous context.

Examination of both correct recall of the target list and incorrect recall from the intervening list provided measures of the relative interference from both lists, as well as the overall levels of recall from both lists. These additional measures were used to separately index the role of context similarity and context retrieval. Beyond the list-length effects on target list recall, between-list tasks that reduced intervening list interference also produced incorrect intervening list recall that was only sensitive to the length of the intervening list. In other words, with recall between the lists, a recalled item was affected only by the length of the study list to which it belonged, regardless of whether that list was the target list or the intervening list. This pattern suggests that sometimes memory retrieval was achieved with a probe context unique to the target list and other times memory retrieval was achieved with a probe

Table 2
Best-Fitting Parameters for Each Experiment

Experiment and task between lists	Parameter			
	M	C_T	R_T	R_I
Experiment 1			1.00	.02
Recall (45 s)	.02	.14		
No test (2.5 s)	.08	.10		
Experiment 2			.25	.10
Recall (60 s)	.03	.17		
Recognition (~8 s)	.44	.07		
Experiment 3			.52	.03
Recall (60 s)	.07	.10		
Letter completion (60 s)	.07	.20		
2-back task (60 s)	.33	.05		

Note. M = proportion of shared context (mixing); C_T = probability of target context retrieval; R_T = recovery for target words; R_I = recovery for intervening words.

context unique to the intervening list. This is a sensible result if the two list contexts are well differentiated, and furthermore, some retrieval attempts fail to reinstate the target list context (i.e., a failure in context retrieval). Furthermore, the data suggest that reinstatement of the target context varied as a function of list isolation. More isolated lists (i.e., dissimilar from the intervening list) required a greater reliance on target list context retrieval to support performance levels comparable to nonisolated lists. This suggests that use of context in retrieval is efficient, if not somewhat lazy, using the current context if it reasonably matches the target list (although this produces interference from the intervening list), but resorting to more effortful reinstatement of the target list context if current context is inadequate.

Beyond the need for context change and context retrieval, the full data pattern also suggested the need for some sort of filtering or censoring process that overall reduced the levels of intrusion from intervening list items. This need is highlighted by examining the relative interference from each list as compared with the relative degree of recall from each list—conditions that failed to promote context change revealed equal amounts of interference from both lists, and yet correct recall from the target list was greater than incorrect recall from the intervening list. To demonstrate the adequacy of these three mechanisms—context similarity, context retrieval, and censoring—we presented a simple multinomial model that included the corresponding three stages of context selection, competitive sampling of items based on the selected context, and finally recovery of sampled items. Censoring was implemented as different recovery rates for the target list as compared with the intervening list, although this could be equivalently implemented as recovery followed by the possibility of rejection for intervening list items. Besides demonstrating that these three mechanisms captured the reported results, producing a computational account allowed quantitative specification for magnitude of each mechanism. In this manner, we determined that changes in context similarity were the predominant effect of between-list task manipulations and that fitting the data by allowing both context similarity and context retrieval to vary as a function of between-list task produced a systematic reciprocal relationship between these mechanisms; situations that produced dissimilar contexts also produced more reliance on target list retrieval.

Implications for Theories of Recall

Theories of free recall data assume that a probe context is used to limit the potential set of recalled memories. However, these theories often fail to define the nature of context or identify factors that affect endogenously generated context. Despite this lack of specification, nearly all theories assume that target list context is fully and properly reinstated and, furthermore, that the reinstated context is unique to the studied list, thus eliminating interference from other lists or preexperimental memories. Our investigations with the list-before-the-last paradigm revealed that this assumption may be false in many situations. Furthermore, the changeable nature of context reinstatement in our results highlights the need for more well-specified theories of context development, context change, and context use.

An initial attempt at specifying context change was developed in Mensink and Raaijmakers's (1988) context fluctuation model. This account assumes that context constantly changes in a seemingly stochastic fashion, such that delay or processing between one encoded memory and a subsequent encoded memory determines the extent of context similarity. Howard and Kahana's (2002) TCM built on this initial context model by assuming that the mechanism that drives this gradual context change is the updating of current context on the basis of the preexperimental context of the encoded memories. Beyond this context change that occurs during study, TCM also assumes that the act of recalling serves to update current context in a discrete and complete fashion because of reinstatement of the encoded context on successful recall. Thus, TCM predicts that the act of recalling will more greatly promote context change. Our results appear to validate this prediction. Although TCM was developed in relation to episodic recall, our results indicate that the theory may apply equally to long-term recall of lexical-semantic information. For instance, Experiment 3 found that the task of filling in a missing letter was comparable to episodic recall in promoting context differentiation. In addition, our results reveal that recall from short-term memory does not promote context change, suggesting that retrieval from short-term memory does not include retrieval of preexperimental context and the associated updating of the temporal context.

Besides TCM, another model that provides a considerable contribution to our understanding of context in episodic memory tasks is the BCDMEM (the bind cue decide model of episodic memory; Dennis & Humphreys, 2001) of recognition memory. Dennis and Humphreys (2001) made the distinction between context-noise models versus item-noise models, with the former using the item to probe memory (thus interference arises from competing retrieved contexts), whereas the latter use context to probe memory (thus interference arises from competing items that match the probe context). BCDMEM stands in apparent opposition to the other global memory models of recognition memory by assuming that retrieval uses the items first, followed by a comparison of the retrieved context (i.e., a context-noise model). This model successfully explains a wide range of findings, including word frequency effects, by assuming that variability in the different contexts associated with an item play an important role in determining interference and forgetting. Although this model is only applicable to recognition memory, our results may help specify the manner in which context changes over time, thus specifying situations that do or do not promote context variability between multiple encounters with the same item.

Other Paradigms That Produce Context Change

Because no exogenous context was provided to differentiate the lists in our experiments, it was assumed that any apparent context changes reflected the process of developing endogenous context. Nevertheless, these results find concordance with manipulations of exogenous context. For instance, testing memory in an environmental context that differs physically from the study context results in lower performance as compared with testing in the original context (e.g., Godden & Baddeley, 1975; see Smith, 1988, for a review). Thus, when explicitly given an exogenous context, it appears that memory retrieval is somewhat lazy and relies on the current context rather than reinstating the original context. Something similar was seen in our results for conditions that produced more intervening list interference. Rather than examining retroac-

tive interference and context change, Dallett and Wilcox (1968) found evidence of proactive interference effects by presenting different lists of words on different days, with half of the participants receiving a change in the environmental context. Although recall became worse with each successive day (i.e., a buildup of proactive interference), a change in the environmental context improved recall (i.e., a release from proactive interference because of the exogenous context change). Analogously, we found that intervening list intrusion rates were unaffected by the target list length for conditions that promoted endogenous changes in context (i.e., reduction in proactive interference).

A paradigm that has many surface similarities to the list-before-the-last paradigm is that of directed forgetting. In a typical directed forgetting experiment, participants are given a list of words that may or may not be followed by the instructions that the just-learned list should be forgotten. This is followed by a second list and then a final test in which participants are asked to recall from both lists, even if they were previously told to forget the first list. The interesting finding is that the forget instructions appear to work, resulting in worse performance on the first list, but at the same time, these instructions result in better performance on the second list, suggesting that there are costs and benefits to directed forgetting (e.g., Bjork, 1970; Reitman, Malin, Bjork, & Higman, 1973). These findings are often attributed to an inhibition process in relation to the to-be-forgotten list (Bjork, 1989), and individual differences in the magnitude of this effect are even related to clinical disorders that involve inhibitory deficits (e.g., Cloitre, 1998; McNally, 2005). An alternative explanation builds on the work of traditional retroactive interference theories and instead appeals to the specificity of the context used to probe memory (Smith, 1988), thereby explicating the observed trade-off between performance on each list (e.g., Sahakyan & Kelley, 2002). Analogously, overall target list performance levels were actually slightly better in our experiments for conditions that failed to promote context change, and at the same time these conditions also produced slightly less intervening list interference. In the multinomial model, this occurred because of the failure to reinstate the target list context; if the two contexts were similar, the current context was sufficient to support prior list recall, but if the contexts were dissimilar, the current context only matched the intervening list. This similarity suggests that the forget instructions in directed forgetting may likewise serve as another manipulation that drives context change. For instance, this may occur if the forget instructions result in a reset of the temporal context, thus clearing out the tendency for the prior list's context to blend into the next list.

Conclusions

Our study finds support for the claim that the act of recalling from long-term memory serves to drive endogenous context change and isolate prior memory episodes. In addition, our results suggest that memory often fails to reinstate the original study list context, defaulting on the current context. Context reinstatement may be adaptive in its application, with situations that promote context differentiation also producing a greater reliance on context reinstatement, and situations that do not promote context differentiation resulting in the good-enough use of the current context. Although these results do not specify the exact nature of endogenous context representations, they greatly expand our understanding for the manner of context use and context change.

References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97,* 548–564.

Bjork, R. A. (1970). Positive forgetting: The noninterference of items intentionally forgotten. *Journal of Verbal Learning and Verbal Behavior, 9,* 255–268.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger, III, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309–330). Hillsdale, NJ: Erlbaum.

Cloitre, M. (1998). Intentional forgetting and clinical disorders. In J. M. Golding & C. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 395–412). Mahwah, NJ: Erlbaum.

Dallett, K., & Wilcox, S. G. (1968). Contextual stimuli and proactive inhibition. *Journal of Experimental Psychology, 78,* 475–480.

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1136–1146.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145–154.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66,* 325–331.

Hintzman, D. L. (1988). Judgments of frequency of recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 923–941.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269–299.

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

Kucera, H., & Francis, W. H. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology, 15,* 215–233.

McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition. *American Journal of Psychology, 43,* 579–588.

McNally, R. J. (2005). Directed forgetting tasks in clinical research. In A. Wenzel & D. C. Rubin (Eds.), *Cognitive methods and their application to clinical research* (pp. 197–212). Washington, DC: American Psychological Association.

Mensink, G.-J., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review, 95,* 434–455.

Murdock, B. B. (1962). The serial position curve of free recall. *Journal of Experimental Psychology, 64,* 482–488.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104,* 839–862.

Osgood, C. E. (1949). An investigation into the causes of retroactive interference. *Journal of Experimental Psychology, 56,* 132–143.

Postman, L., & Philips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology, 17,* 132–138.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27,* 431–452.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

Reitman, W., Malin, J. T., Bjork, R. A., & Higman, B. (1973). Strategy control and directed forgetting. *Journal of Verbal Learning and Verbal Behavior, 12,* 140–149.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology, 92,* 365–372.

Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition, 24,* 188–201.

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition, 22,* 511–524.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1064–1072.

Shiffrin, R. M. (1970, June 26). Forgetting: Trace erosion or retrieval failure? *Science, 168,* 1601–1603.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Sirotin, Y. B., Kimball, D. R., & Kahana, M. J. (2005). Going beyond a single list: Modeling the effects of prior experience on episodic free recall. *Psychonomic Bulletin & Review, 12,* 787–805.

Smith, S. M. (1988). Environmental context-dependent memory. In G. M. Davies & D. M. Thomson (Eds.), *Memory in context: Context in memory* (pp. 13–34). Chichester, England: Wiley.

Ward, G., & Tan, L. (2004). The effect of the length of to-be-remembered lists and intervening lists on free recall: A reexamination using overt rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1196–1210.

Appendix A

Experiment 2 Recognition Results

The descriptive statistics are shown in Table A1. There was a four-way interaction between type of tasks between the lists, type of foils, length of target list, and length of intervening list, $F(1, 99) = 5.39$, $MSE = .05$, $\eta_p^2 = .05$. Table A2 reports the 2×2 analysis of variance results for recognition broken down according to recall between the lists or recognition between the lists.

Table A1
Experiment 2 Forced-Choice Recognition Performance

	Recall between lists				Recognition between lists			
	Intervening foils		New foils		Intervening foils		New foils	
Intervening list length	Long target list	Short target list	Long target list	Short target list	Long target list	Short target list	Long target list	Short target list
Long	.55 (.03)	.62 (.03)	.67 (.03)	.72 (.02)	.51 (.03)	.50 (.02)	.66 (.03)	.67 (.03)
Short	.66 (.03)	.65 (.03)	.69 (.03)	.78 (.02)	.66 (.03)	.69 (.03)	.78 (.02)	.75 (.03)

Note. Standard errors of the mean are in parentheses.

Table A2
Experiment 2 Forced-Choice Recognition Analysis of Variance Results

	Intervening foils				New word foils			
Between-list task	$F(1, 99)$	MSE	p	η^2	$F(1, 99)$	MSE	p	η^2
Recall								
T	1.25	.07	.27		9.91	.05	<.01	.09
I	7.14	.07	<.01	.07	2.47	.05	.12	
T × I	2.42	.06	.12		<1			
Recognition								
T	<1				<1			
I	48.33	.06	<.001	.33	19.32	.05	<.001	.16
T × I	<1				<1			

Note. Eta squared (effect size) is reported only when the F value was significant. T = target list length; I = intervening list length.

Appendix B

Experiment 3 Between-List Task Performance

The descriptive statistics of the letter completion and 2-back tasks from Experiment 3 are shown in Table B1. Table B2 shows the analysis of variance results for each of these tasks as a function of the list-length combinations.

Table B1
Accuracy in the Letter Completion and 2-Back Tasks in Experiment 3 as a Function of List Length

	Letter completion between lists		2-back between lists	
Intervening list length	Long target lists	Short target lists	Long target lists	Short target lists
Long	.90 (.011)	.97 (.003)	.76 (.022)	.74 (.023)
Short	.95 (.004)	.91 (.008)	.74 (.022)	.74 (.023)

Note. Standard errors of the mean are in parentheses.

Table B2
Results From Analyses of Variance of Letter Completion and 2-Back Tasks in Experiment 3

	Letter completion task				2-back task			
Effect	$F(1, 154)$	*MSE*	*p*	η^2	$F(1, 154)$	*MSE*	*p*	η^2
T	10.48	.01	$<.01$	.06	<1			
I	1.63	.01	.20		<1			
T × I	49.41	.01	$<.001$	.24	<1			

Note. Eta squared (effect size) is reported only when the *F* value was significant. T = Target list length; I = Intervening list length.

Received May 23, 2007
Revision received August 23, 2007
Accepted August 24, 2007 ■