Computational Brain & Behavior (2021) 4:178–190
https://doi.org/10.1007/s42113-020-00090-y

ORIGINAL PAPER

The Moderating Role of Feedback on Forgetting in Item Recognition

Aslı Kılıç[1] · Jessica Fontaine[2] · Kenneth J. Malmberg[3] · Amy H. Criss[2]

Published online: 9 September 2020

Abstract

We conducted three experiments specifically designed to simultaneously evaluate the effects on recognition accuracy of adding items during study and adding items during test. The recognition memory list-length effect (LLE) is small and unreliable (Annis et al. 2015; Dennis et al. 2008), but additional test trials produce a robust decrease in accuracy, termed output interference (OI; Criss et al. 2011; Kılıç et al. 2017). This is puzzling; why should the size of the effect of additional stimulus exposures depend on whether the item was studied or tested (Malmberg et al. 2012)? We found a decrease in accuracy when stimulus exposures were added at any stage. However, the harm of adding items during study was less than the output interference that resulted from testing. In addition, feedback presented during test served as a moderator. When feedback was given, OI was diminished, and the LLE increased. Within the framework of our model, this suggests that testing with no feedback often results in the encoding of additional information in a trace originally encoded during study, and testing with feedback decreases the tendency to update traces during test. Several possible accounts of feedback reducing trace updating are discussed.

Keywords Item recognition · Output interference · List length effect · Memory models · Feedback

Errors in memory range from a daily annoyance to a threat to health and freedom. Understanding the nature of memory errors and the sources of forgetting has important practical and theoretical implications. Most theories of memory assume that forgetting is the result of interference caused by irrelevant memories during retrieval (Anderson et al. 1998; Dennis and Humphreys 2001; Murdock 1982; Raaijmakers and Shiffrin 1981; Reder et al. 2000). Interference is often investigated using a study-test procedure, whereby subjects study lists of items and later memory for those items is tested. Interference is produced when items are encoded during study. For example, when associations are created between similar items (e.g., study AC following the study of AD), making it more difficult to retrieve either pair (Crowder 1976, for a review). Likewise, interference may be a result of storing new memories during testing (Wickens 1970).

Two experimental findings in the recognition literature—the list length effect and output interference—demonstrate increased forgetting with increases in interference from storing additional events in memory during study and test, respectively. We conducted three experiments specifically designed to tease apart the effects of adding items during study and adding items during test on recognition memory, and the results are interpreted within a retrieving effectively from memory model (Shiffrin and Steyvers 1997; Criss et al. 2011). To foreshadow, we find that testing conditions determine the degree to which interference from traces stored during study and test impact performance.

The List Length Effect

The list length effect is a reduction in accuracy that results from increasing the size of the to-be-remembered memory set. There are many demonstrations of this list length effect (LLE) beginning with Strong Jr 1912, see also Cary and Reder 2003; Criss and Shiffrin 2004; Gronlund and Elam 1994; Nobel and Shiffrin 2001). Accordingly, most models of memory assume that representations of other items are a source of

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42113-020-00090-y) contains supplementary material, which is available to authorized users.

✉ Aslı Kılıç
askilic@metu.edu.tr

[1] Middle East Technical University, Ankara, Turkey

[2] Syracuse University, Syracuse, NY, USA

[3] University of South Florida, Tampa, FL, USA

interference and they predict that increasing the number of items studied should harm performance. However, in practice, list length manipulation is susceptible to several confounds, and therefore, surprisingly few recognition experiments have been reported with clear results.

For instance, Dennis and Humphreys identified confounding details of many experimental designs that may cause an artifactual LLE, and they reported null LLEs for recognition when these design elements were controlled (Dennis and Humphreys 2001; Dennis et al. 2008; Kinnell and Dennis 2011, cf. Cary and Reder 2003).

Even when conditions are well controlled, however, empirically establishing the LLE as either present or absent has been challenging because the LLE is predicted to be quite small and recognition memory data are inherently noisy (Shiffrin and Steyvers 1997). For instance, the data from the Dennis et al.'s (2008) LLE experiment in which various confounds were rigorously controlled provided more support for the models that predicted an LLE than those that did not (Annis et al. 2015). Annis et al. simulated data from a context noise model, BCDMEM (Dennis and Humphreys 2001), in which the LLE is not predicted, and from an item noise model, retrieving effectively from memory (REM) (Shiffrin and Steyvers 1997), which predicts LLE. The simulated data was found to be not diagnostic in distinguishing between the predictions of these models.

Setting these issues aside for the moment, there are two unambiguous empirical facts: The magnitude of the LLE is rather small, sometimes so small so as to not be measurably different from zero. When the LLE is present in yes-no recognition performance, it takes the form of a mirror effect—hit rates are greater and false alarm rates are lower for a shorter list compared with a longer list (e.g., see Gronlund and Elam 1994; Cary and Reder 2003). This suggests that the interference due to encoding additional traces prior to recognition memory testing is not usually a large source of forgetting. Further, the long-standing debate over the existence and magnitude of the recognition memory LLE suggests that there may be some unidentified factors that increase or reduce the magnitude of the LLE.

Output Interference

Output interference (OI) is the finding that episodic recognition accuracy decreases with an increase in the number of test trials (Criss et al. 2011; Kılıç et al. 2017; Koop et al. 2015; Murdock and Anderson 1975; Ratcliff and Hockley 1980; Roediger and Schmidt 1980). In contrast to the small and noisy LLE, OI is relatively larger in magnitude and highly reliable. Moreover, OI is primarily reflected by a decrease in hit rates across test trials and a smaller, variable pattern in the false-alarm rates in contrast to the mirror pattern of the LLE. Together these two findings appear to present a paradox: why is recognition memory minimally harmed by increasing the number of to-be-learned items but increasing the number of test items substantially decreases accuracy and why do the patterns of decreasing accuracy differ?

These two effects are typically reported in different experimental paradigms with different constraints. List length experiments often have a variable delay to equate the duration between the first study trial and the first test trial, a short test list to equate test length for short and long study lists, and an encoding task to minimize differential attention across the study lists. OI experiments often include feedback during test to control motivation across the test trials. Therefore, one possible explanation for this apparent LLE-OI paradox is simply that the experimental designs foster different outcomes. We addressed this possibility by evaluating both the LLE and OI simultaneously each within a constant experimental design.

REM Model

Another account of the LLE-OI paradox is that different encoding operations underlie the two phenomena, producing both the quantitative and qualitative differences in observed behavior.[1] The encoding operations we focus on are implemented in the REM model framework (Shiffrin and Steyvers 1997; Malmberg et al. 2004; Criss 2006; Criss et al. 2011; Kılıç et al. 2017). REM assumes that items are composed of features. Some feature values are relatively rare, and thus highly diagnostic, and others are relatively common and shared by many different words. The degree to which items share features determines how similar their representations are and hence their confusability.

REM assumes episodic memory traces stored during an event (one per stimulus in a typical laboratory situation) are a noisy and incomplete representation of the items that were studied. Some features are not stored during study and other features are stored incorrectly, due to limited study time, attention, orienting tasks, strategic encoding, or errors in encoding. At test, the to-be-remembered context is activated and the memory traces in that context become available for comparison with the memory probe. The features of the memory probe are compared with each of the episodic traces and the degree of match between the test stimulus and each memory trace is computed. A recognition decision is based on the global match of the probe to the memory traces (i.e., the average of the match value between the probe and each stored trace). As the number of activated traces in episodic memory increases, the number of spurious matches increases due to random matching of common features in non-target traces. Additional item noise resulting from increases in the number of spurious matches reduces the hit rate and increases the false

[1] Here we present a simplified explanation of REM to illustrate the key factors related to item noise interference and specifically the LLE and OI.

alarm rate, producing a mirror-patterned list-length effect. This item noise is the source of the LLE and one source of item-based interference.

Updating Stored Memory Traces

Updating of memory traces is the second mechanism that produces item noise interference, and the profound effects on recognition performance are qualitatively different than the effect of adding an additional trace according to REM. Repetition during study leads to updating a memory trace by storing features that were left unstored during the initial encoding resulting a more complete memory trace (Malmberg et al. 2004). Updating a memory trace causes a reduction in the noise associated with that trace—it is a better match to the target from which it was generated and it is a worse match to any other item (see Criss 2006, 2009, 2010; Criss and Koop 2015; Kılıç et al. 2017). That is, an updated trace is differentiated. Differentiation models were developed to account for data that extant global-matching models could not (REM and McClelland and Chappell 1998, see Criss and McClelland 2006) including the null list strength effect (Ratcliff et al. 1990; Shiffrin et al. 1990).

Repetitions also occur when a target item is tested. Criss et al. (2011) introduced encoding during test to the REM framework to account for output interference. Specifically, if the item is recognized, the best matching episodic memory trace is updated.[2] Otherwise, a new trace is stored. If an item is judged to be new, then a new trace is stored. Thus, at study and at test, repeated items are sometimes stored in new memory traces and sometimes cause updating of a previously stored memory trace. Updating a memory trace differentiates it from other traces, reducing its confusability by increasing the number of mismatching features between the trace and a different item and increasing the number of matching features between the trace and its corresponding target.

Figure 1 shows that updating memory traces versus storing new traces produce fundamentally different patterns of data. This simulation begins with a fixed number of study items and tracks the performance across a test list containing half foils and half targets. In the 'new-trace' model, an additional memory trace is stored on each test trial no matter. Hence, the number of traces in the relevant memory set is directly related to the number of trials, and as they increase in number, the HR decreases and FAR increases. In other words, storing new traces hides the signal in additional noise making it more difficult to discriminate between targets and foils. In contrast, the 'update model' assumes traces are updated when the test item is judged to be old; otherwise, a new trace is stored (i.e., the Criss et al. (2011) model). Updating traces at test produces

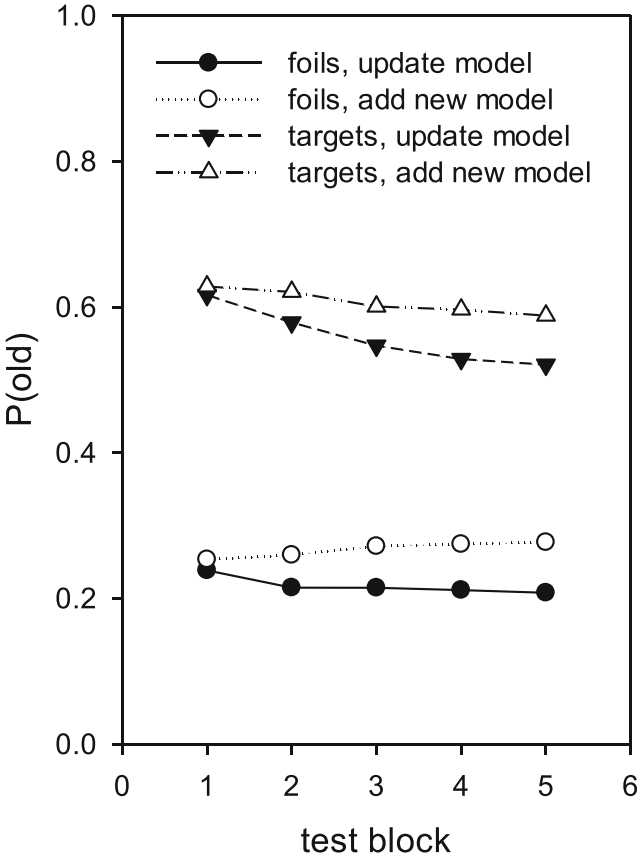

Fig. 1 REM simulations of a model where a new trace is added for every test trial ("add new model") compared with a model where a new trace is added when the item is judged to be new and the best matching trace is updated when a test item is remembered ("update model"). Adding a new trace on each trial results in a mirror pattern with hits decreasing and false alarms increasing. Updating results in differentiation and a steep decrease in hit rate and a slight decrease or flat false alarm rate

differentiation and substantially decreases the HR and slightly decreases the FAR *for subsequently tested items*. Note some endorsed items are foils (i.e., false alarm). In that case, endorsing a foil causes an update of a trace which can be tested subsequently. According to this model, the most highly activated trace in response to the probe is updated. Sometimes, the updated trace represents the test item (i.e., a hit), but sometimes, a trace representing a different item is spuriously updated with features of the current test item. This could occur if a false alarm is made. Consequently, HR and FAR decreases because fewer random matches occur between the subsequent test probes and the updated, well-encoded memory traces. That is, the better stored any memory trace, the less likely it is to match *other* items. Note that the magnitude of decrease in HR is more substantial when traces are updated than when new traces are stored and the effect on FAR is in different directions (increasing with new traces and decreasing or remaining flat with updating).

In REM, the LLE has been modeled with a new-trace mechanism and OI has been modeled with an update

[2] Note that this could result in the updating of the trace representing a different target, although this occurs infrequently when the items are randomly similar.

mechanism. These two different approaches have successfully accounted for empirical pattern of results in separate studies. However, these two effects have not been simultaneously measured. Further the methodological concerns in the two paradigms have resulted in different experimental designs. The goal of this paper is to simultaneously evaluate the impact of adding items during study and test under identical conditions. The empirical effects are of interest on their own but are particularly interesting in that they can help constrain models of memory, specifically the two assumptions described above—adding new traces compared with updating traces.

Experiment 1

The primary experimental manipulations in this and the forthcoming experiments are study-list length and the number of test trials. (Though, this is not manipulated per se, but rather, performance is plotted as a function of test trial.) Secondary manipulations were developed to address specific methodological concerns. For instance, a delay between study and test was included as is common in many studies of the LLE. In addition, feedback was also provided during test for half of the conditions to address the classic hypothesis that recognition OI may be due to the decline in motivation, attention (or likewise an increase in boredom) across the test list (but see Criss et al. 2017). By giving feedback, we intended to simply encourage participants to continue to try on each trial so as to avoid unpleasant negative feedback.

Methods

Participants

Participants in the experiment were 288 undergraduates from the Syracuse University research participation pool who received partial course credit. One participant who performed much lower than chance (A' was less than 0.40 and a chance value of A' is 0.5) was removed from the subsequent analysis.

Materials

The word pool consisted of 800 high frequency words between 4 and 7 letters in length and ranging between 9 and 13 log frequency ($M = 10.46$) in the Hyperspace Analog to Language corpus (Balota et al. 2007). Words were randomly assigned to condition for each participant.

Design and Procedure

Participants were randomly assigned to a condition in the 2 (feedback) × 2 (study list length) × 2 (delay) design by scheduling block (everyone in a block, N between 1 and 10, participated in the same condition). Participants received a study list with each item presented in the middle of the screen for 1 s with a 100-ms blank screen separating trials. The study list was either 75 items (short list) or 200 items (long list). Following study, a 45-s distractor addition task was completed, followed by a delay of 0 or 15 min. The delay was filled with a puzzle activity.

The test list consisted of 150 trials of self-paced single item recognition trials. The test list included 75 targets (the last 75 that were studied in both conditions in order to equate study-test lag) and 75 foils, randomly intermixed. Participants judged (yes or no) whether the test word was studied. Feedback (correct or wrong) was provided for 100 ms following each response, or a blank screen (no feedback condition) was presented for the same duration.

Results and Discussion

We first conducted an analysis of A' to highlight the critical factors before turning to an analysis including OI (Table 1).[3] Sensitivity significantly decreased with delay, $F(1,279) = 21.12$, $p < .001$, $\eta_p^2 = .07$, and longer study lists $F(1,279) = 18.70$, $p < .001$, $\eta_p^2 = .06$. There was an interaction between feedback and list length, $F(1,279) = 3.91$, $p = .05$, $\eta_p^2 = .01$. Post hoc comparisons using t test with Bonferroni adjustment showed that the LLE was effective when feedback was provided, $t = 4.72$, $p_{Bonf} < .001$, but not when feedback was not provided, $t = 1.59$, $p_{Bonf} = 0.68$. No other main effects or interactions were significant.

To evaluate the pattern of output interference, the test was divided into 5 blocks of 30 trials, resulting in a mixed design with test block as the within-subject factor and feedback, list length, and delay as between-subject factors (Fig. 2).[4] FAR significantly increased with a delay between study and test, $F(1, 279) = 9.72$, $p = .002$, $\eta_p^2 = .03$, producing at least some of the change in sensitivity described above. Main effects of test block, $F(4,1116) = 5.27$, $p < .001$, $\eta_p^2 = .02$ and feedback, $F(1,279) = 20.50$, $p < .001$, $\eta_p^2 = .07$ were qualified by interactions. Both manipulations of primary interest (test block and study list length) interacted with feedback, $F(4,1116) = 2.96$, $p = .02$, $\eta_p^2 = .01$, and $F(1,279) = 4.18$, $p = .04$, $\eta_p^2 = .02$, respectively. When feedback was provided, FAR increased across test block. The difference in FAR between the first test block ($M = .29$, $SD = .19$) was significantly lower than the FAR in the last block ($M = .34$, $SD = .19$ as post hoc comparisons measured by t test with Bonferroni adjustments showed, $t = 5.78$, $p_{Bonf} < .001$, but FAR remained

[3] Wherever A' is reported, we also conducted analyses with d' and the pattern of data was the same.

[4] Throughout we conducted analyses with different block sizes to investigate the robustness of the effect. The patterns of data hold regardless of block size.

Table 1 Discriminability in each condition of each experiment. Experiments 1 and 2 used yes/no single item recognition therefore A′ is a suitable measure of the ability to discriminate between targets and foils. Experiment 3 used 2 alternative forced choice and we report the proportion of correct trials on which the target was identified (P(c))

Delay	Feedback	List Length	Experiment 1 A′	Experiment 2 A′	Experiment 3 P(c)
0 min	Yes	Short	.74	.76	.70
		Long	.67	.68	.62
	No	Short	.73	.74	.67
		Long	.70	.71	.61
15 min	Yes	Short	.70		.65
		Long	.63		.58
	No	Short	.67		.61
		Long	.64		.60

fairly flat in the baseline condition, $t = 0.70$, $p_{Bonf} = 1$. This finding is consistent with a model in which an existing trace was updated with erroneous information during the commission of a false alarm. The list length by feedback interaction was further analyzed by post hoc comparisons. When feedback was not provided, LLE was not observed, $t = 0.134$, $p_{Bonf} = 1$. However, FAR was greater in the long condition ($M = .44$, $SE = .02$) then in the short condition ($M = .36$, $SE = .02$) when feedback was provided, $t = 2.93$, $p_{Bonf} = .02$. In other words, feedback produces performance that is consistent with a model where a new trace is stored on each trial rather than updating an existing trace with new information.

HR was higher for short compared with long study lists, $F(1,279) = 6.87$, $p = .009$, $\eta_p^2 = .02$. The other main effects, test block $F(4,1116) = 23.63$, $p < .001$, $\eta_p^2 = .08$ and feedback $F(1,279) = 28.09$, $p < .001$, $\eta_p^2 = .09$, were qualified by interactions. Even though the main effect of delay was not significant, $F(1,279) = 0.74$, $p = .39$, $\eta_p^2 = .003$, delay interacted with test block such that the decrease in HR across test blocks was reduced after a delay compared with testing with no delay as revealed by a significant interaction, $F(4,1116) = 7.51$, $p < .001$, $\eta_p^2 = .03$. This finding is likely due to approaching floor performance when the test is delayed. Post hoc comparisons measured by t tests with Bonferroni adjustments showed that when there was no delay, HR in test block 1($M = 0.67$, $SD = 0.15$) was greater than HR in test block 5 ($M = 0.53$, $SD = 0.21$), $t = 8.38$, $p_{Bonf} < 0.001$, suggesting evidence for the effect of test block. When a delay was introduced, HR in test block 1 ($M = 0.57$, $SD = 0.20$) did not differ significantly from HR in test block 5 ($M = 0.53$, $SD = 0.21$), $t = 2.57$, $p_{Bonf} = .44$. Similarly, HR in the last block did not differ across delay conditions, $t = 0.68$, $p_{Bonf} = 1$. The decrease in HR across

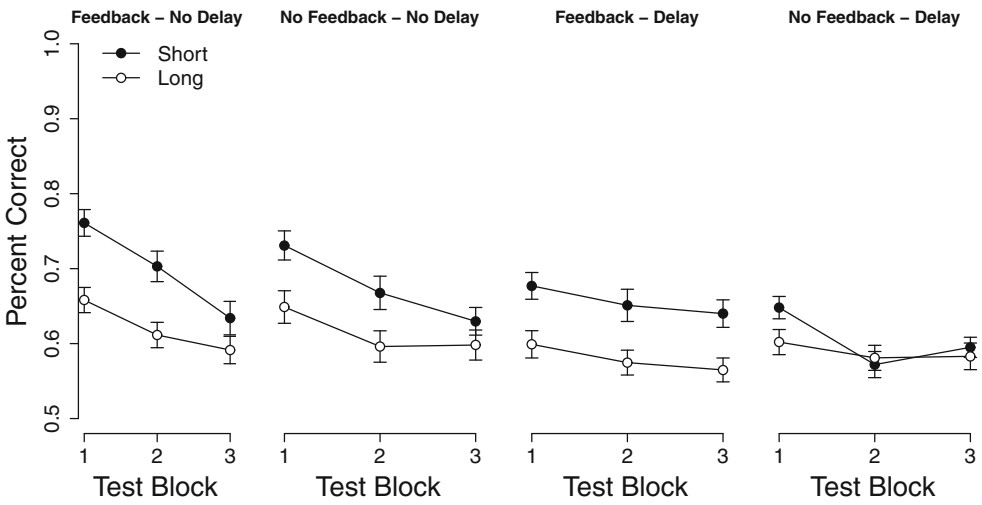

Fig. 2 Probability of calling a test item old in experiment 1. Black symbols are targets and white symbols are foils. Squares represent the long study list, and circles represent the short study list conditions. Error bars represent standard error of the mean

test blocks was also reduced in the presence of feedback during test, $F(4,1116) = 5.37$, $p < .001$, $\eta_p^2 = .03$. Post hoc t tests with Bonferroni adjustments showed that when feedback was provided, HR in test block 1 ($M = 0.63$, $SD = 0.15$) did not differ significantly from HR in test block 5 ($M = 0.59$, $SD = 0.19$), $t = 2.78$, $p_{Bonf} = .26$. However, in the absence of feedback, HR in block 1 ($M = 0.60$, $SD = 0.22$) dropped in block 5 ($M = 0.46$, $SD = 0.22$), $t = 8.22$, $p_{Bonf} < .001$. This pattern is again consistent with the idea that feedback elicits behavior more consistent with a model where a new trace is added on each test trial.

In summary, both OI and a LLE were observed under constant experimental methodology. OI was relatively large in magnitude (compared with the size of the LLE), consistent with the REM model where adding new traces during study results in the storage of a new trace but test trials sometimes results in the storage of a new trace and other times results in the updating of a stored memory trace. This differentiation of traces during test causes the updated traces to become even less similar to other items (e.g., those tested later during the test) reducing both the HR at a faster rate than under a model where a new trace is stored and reducing the FAR.

We included feedback in the current study to maintain motivation throughout the test with an aim to preclude the ambiguity related to changing vigilance. Previously, many studies have shown that feedback has virtually no impact on accuracy in recognition memory (Han and Dobbins 2008, experiment 1; Kantner and Lindsay 2010; Criss et al. 2011) though a few have shown that feedback may affect response bias (Kantner and Lindsay 2010; Koop et al. 2015; Starns et al. 2010). The results in the current experiment showed a moderating effect of feedback on LLE and OI, especially observed in HR and FAR without a strong effect on accuracy. One hypothesis for this finding is that the feedback effects reflect changing criteria between the different conditions, especially given that list length and feedback are between-subject manipulations in this experiment. We evaluate this possibility by conducting the feedback and list length conditions in a within participant design (experiment 2) and applying a forced choice testing paradigm (experiment 3).

Experiment 2

The purpose of this experiment was to replicate the OI and LLE findings and the role of feedback as a moderator. There were a few changes to the design, the most important of which was that the manipulation of study list length and feedback were within-subject and the delay manipulation was eliminated. We also considered the possibility if feedback itself was not critical, but rather, any simple sensory stimulation following the decision might cause an additional trace of the test probe to be stored. To address this, we present the word 'next' following each response in the no feedback condition. This equates the conditions in a sense that a word is visually presented for the same duration after each response, isolating the role of actual feedback on performance.

Methods

Participants

Participants in the experiment were 41 undergraduates from the Syracuse University research participation pool who received partial course credit, 1 of whom was excluded due to technical problems.

Materials

The word pool consisted of 2929 words between 4 and 8 letters in length and of those 2912 in the Hyperspace Analog to Language corpus (Balota et al. 2007), ranged between 1 and 14 log word frequency ($M = 8.71$). Words were randomly assigned to condition for each participant and did not repeat (except when presented as a target during a test).

Design and Procedure

Participants completed 8 rounds of study test, 4 during each of 2 days. The days were separated by approximately 1 week. Each day, the 4 rounds included each condition of the 2 (feedback at test) × 2 (study list length) within-subject design, randomly ordered. In the no feedback conditions, the word 'next' was presented for the same duration as the feedback. All other details were identical to the experiment 1 no delay conditions.

Results and Discussion

The results replicate that A' is better for shorter than longer study lists, $F(1,39) = 29.73$, $p < .001$, $\eta_p^2 = .43$. We should note that when list length conditions are countered balanced in a within subject design, the order of the study list length might obscure the LLE such that when long list is studied before short list, LLE disappears (Brandt et al. 2019; Fox et al. 2020). However, the effect size of the list length variable was reasonably high, suggesting that such a possibility is very unlikely in the current experiment. LLE interacts with feedback, $F(1,39) = 4.85$, $p = .03$, $\eta_p^2 = .11$. Post hoc comparison using t test with Bonferroni adjustments showed that the LLE is only present when feedback is provided during test $t = 5.09$, $p_{Bonf} < .001$, but not in the absence of feedback $t = 1.62$, $p_{Bonf} = 0.66$ (Table 1).

OI was evaluated by dividing the test into 5 equally sized blocks, each of which contained 30 test trials. A 2 (study list length) × 2 (feedback) × 5 (test block) repeated measures

ANOVA was conducted for false alarms and hits. The data are plotted in Fig. 3. FAR was higher for longer lists, $F(1,39) = 8.89$, $p = .005$, $\eta_p^2 = .19$, and when feedback was provided $F(1,39) = 26.57$, $p < .001$, $\eta_p^2 = .41$. FAR slightly increased across test block, $F(4,156) = 5.38$, $p = .015$, $\eta_p^2 = .08$. Contrasts showed that mean FAR increased from test block 1 ($M = .32$, $SD = .14$) to test block 5 ($M = .34$, $SD = .15$), $t(156) = 2.39$, $p = 0.02$. Neither of the interactions across list length, feedback, and test block variables exceeded the significance criterion value.

HR was lower when feedback was withheld ($M = .62$, $SD = .15$) compared with being provided ($M = .65$, $SD = .13$), $F(1,39) = 8.44$, $p = .006$, $\eta_p^2 = .18$, and for long lists ($M = .62$, $SD = .14$) than for short lists ($M = .65$, $SD = .14$), $F(1,39) = 9.19$, $p = .004$, $\eta_p^2 = .19$. HR decreased across test block, $F(4,156) = 34.30$, $p < .001$, $\eta_p^2 = .47$. The decrease in HR across test block interacted with feedback showing a greater decline in the absence of feedback, $F(4,156) = 4.77$, $p < .001$, $\eta_p^2 = .11$. Post hoc comparisons measured with t test with Bonferroni adjustments revealed an absence of a significant difference across feedback and no feedback conditions at the beginning of the test ($t = 0.182$, $p_{Bonf} = 1$), while mean HR was found to be greater in the final test block for the feedback condition ($M = .62$, $SD = .13$) compared with the mean HR in the no feedback condition ($M = .53$, $SD = .14$), $t = 4.71$, $p_{Bonf} < .001$.

The interaction between list length and feedback was significant for HR, $F(1,39) = 4.11$, $p = .05$, $\eta_p^2 = .10$. Specifically, HR was greater for short lists ($M = .68$, $SD = .13$) compared with long lists ($M = .63$, $SD = .13$) when feedback was provided as post hoc analysis with Bonferroni adjustments showed, $t = 3.5$, $p_{Bonf} < .003$. This effect disappeared when feedback was not provided, $t = 0.69$, $p_{Bonf} = 1$. Together, these findings suggest that the combined effect on discriminability (A') replicated the finding that the difference in accuracy between short and long study lists was magnified when feedback was provided during test.

In summary, OI and LLE were simultaneously observed replicating experiment 1. These two effects attenuated when list length and feedback was respectively manipulated within subjects as was manipulated between subjects in experiment 1. These findings suggest the possibility that receiving feedback during test increases the tendency to add new traces in memory during the commission of a false alarm, increasing the noise caused by other memory traces, whereas a lack of feedback increases the tendency to erroneously modifying an existing trace further reducing recognition accuracy via a decrease in HR. The next experiment aims at evaluating whether this moderating effect of feedback would be observed when response bias is controlled in a forced choice recognition task.

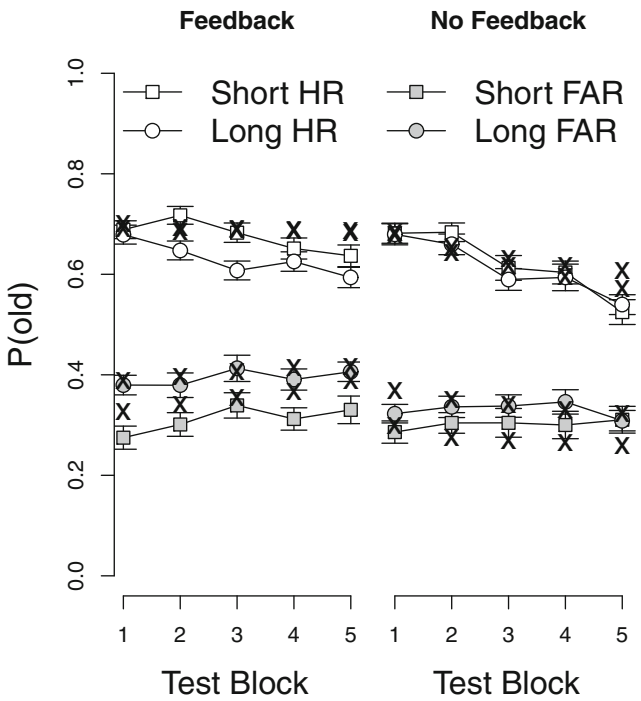

Fig. 3 Probability of calling a test item old in experiment 2. White symbols are targets and black symbols are foils. Circles represent the long study list, and squares represent the short study list conditions. Error bars represent one standard error. The model predictions are represented as x

Experiment 3

The primary goal of this experiment was to replicate the findings we have observed in experiments 1 and 2 and establish whether the effects are dependent on a flexible response bias. Specifically, we were interested in the LLE and OI and the aforementioned role of feedback in the LLE and magnitude of OI. The only change in methodology was the use of two-alternative forced choice (2AFC) for testing. 2AFC has the advantage of being resilient against criterion shifts. The use of 2AFC is particularly useful for evaluating whether the moderation of the LLE and OI by feedback is due to a shift in the criterion. Most accounts of 2AFC assume that decision-maker directly compares the familiarity of the two test items to one another and selects the item with the highest mnemonic evidence (Swets and Green 1961; Malmberg and Murnane 2002; Criss et al. 2011). Hence, 2AFC is criteria free.

Methods

Participants

Participants in the experiment were 326 undergraduates from the Syracuse University research participation pool who

received partial course credit. An additional 5 students participated but were excluded due to technical errors. Data from 3 participants were excluded from the subsequent analysis due to below chance performance. More specifically, performance that is lower than 0.4 in a 2AFC task indicates that the participant used the incorrect key mappings such that to select the word on the left they hit the right key and vice versa. Therefore, data from those participants were excluded and the remaining sample size was 323.

Materials

The Word Pool Was the Same as That Used in Experiment 1

Design and Procedure

The design was identical to experiment 1 with two exceptions. First, rather than presenting 75 targets and 75 foils randomly intermixed as single items during test, the targets and foils were randomly paired and one of each was presented as a pair for a 2 alternative forced choice test. Participants indicated which of the two items was studied. Second, rather than a blank screen separating trials in the no feedback condition, the word 'next' appeared centered on the screen as in experiment 2.

Results and Discussion

To evaluate output interference the test was divided into 3 blocks of 25 trials each, resulting in a mixed design with test block as the within-subject factor and feedback, list length, and delay as between-subject factors (Fig. 4). The results largely replicate experiment 1 and experiment 2. Namely, accuracy decreased with delay, $F(1,315) = 21.26$, $p < .001$, $\eta_p^2 = .06$, study list length, $F(1, 315) = 35.45$, $p < .001$, $\eta_p^2 = .09$, and test block, $F(2,630) = 36.204$, $p < .001$, $\eta_p^2 = .10$. The LLE was enhanced when participants were given feedback as revealed by an interaction between list length and feedback, $F(1, 315) = 3.87$, $p = .05$, $\eta_p^2 = .01$. For the participants who studied long lists, accuracy was not affected by feedback as post hoc comparisons suggest, $t = 0.11$, $p_{Bonf} = 1$. The percentage of correctly selecting the target was similar across feedback ($M = .60$, $SD = .49$) and no feedback conditions ($M = .60$, $SD = .49$). On the other hand, for the participants who studied short lists, receiving feedback affected the percentage of correctly selecting the target, $t = 2.65$, $p_{Bonf} = .05$, such that the percentage was greater when participants received feedback ($M = .67$, $SD = .48$) than when they did not ($M = .64$, $SD = .47$).

Additionally, there was an interaction between test blocks and delay, which showed that the magnitude of OI was dampened when test was delayed, $F(2, 630) = 5.88$, $p = .003$, $\eta_p^2 = .02$, which could again be likely due to the overall decrease in accuracy with delay compressing the scale. Post hoc comparisons measure by t test with Bonferroni adjustments showed that at the beginning of the test block, participants were more accurate when the test list was presented right after the study list ($M = .70$, $SD = .46$) than being presented after a delay ($M = .63$, $SD = .48$), $t = 5.22$, $p_{Bonf} < .001$. However, this pattern disappeared towards the end of the test list, such that the accuracy levels were similar in the delay ($M = .61$, $SD = .48$) and no delay ($M = .59$, $SD = .49$) conditions, $t = 1.34$, $p_{Bonf} = 1$.

Finally, an interaction between test block and list length showed that OI was dampened when items were studied in a long list, $F(2, 630) = 3.06$, $p = .05$, $\eta_p^2 = .01$, which could also be due to an overall decrease in accuracy in long lists. For example, the average rate of indicating the target word correctly at the beginning of the test preceding a short list ($M = .70$, $SD = .46$) was greater than that of the test following a long list ($M = .62$, $SD = .48$), as post hoc comparisons showed, $t = 5.9$, $p_{Bonf} < .001$. Towards the end of the test list, the percentage of correctly selecting the target word decreased with different rates for long and short conditions. For the long list condition, performance reached its asymptote at the second block ($M = .59$, $SD = .49$), as revealed by a lack of a significant difference between the second and the third block ($M = .58$, $SD = .49$), $t = 0.006$, $p_{Bonf} = .61$. That might be due to the reason that participants started with a lower performance at the beginning of the test and their accuracy dropped from there until reaching an asymptote.

In summary, the observations of a decrease in accuracy with more study and more test items and larger such effects with feedback during test were replicated under 2AFC testing. This suggests that these effects are rooted in the interference in accurately retrieving from memory rather than meta-cognitive affects in the decision process.

REM Model

We describe the predictions that the model makes for OI and LLE based on many prior papers. However, as we note, these two empirical phenomena have not been studied simultaneously. To ensure that the model is behaving as anticipated, we fit the model initially to experiment 2. That is mainly because in experiment 2, there was no delay manipulation and therefore we aimed at simulating the data from the most basic experiment, which addresses the direct relationship between OI and LLE and the moderating effect of feedback on LLE. In the supplementary material, we present the model fits of experiments 1 and 3 only for the no delay conditions.

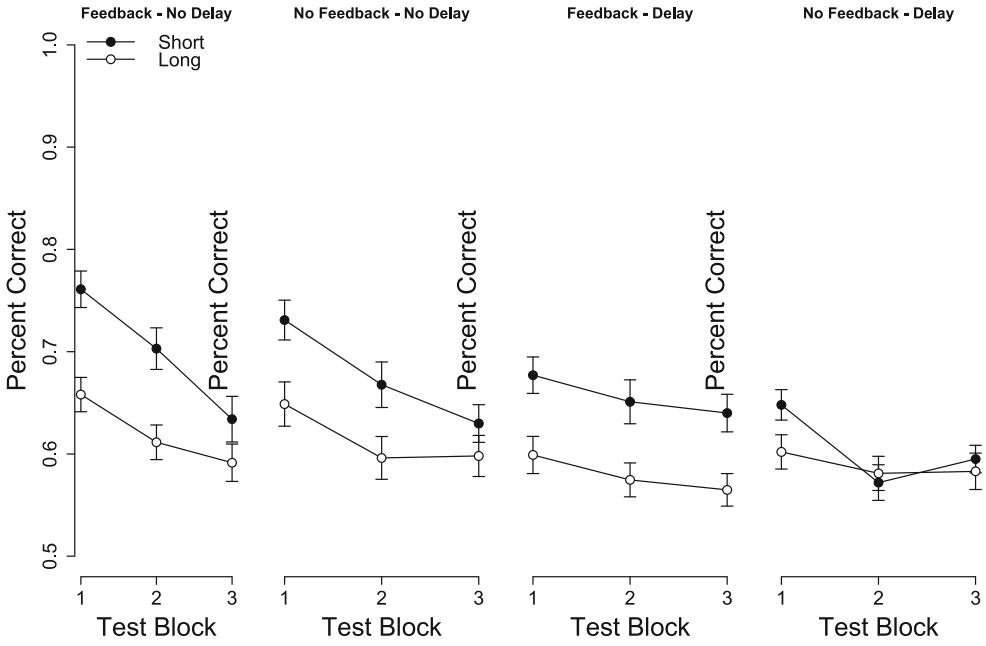

Fig. 4 Percent correct in the 2-alternative forced choice task of experiment 3. Black symbols represent the short list and white symbols the long study list. Error bars are +/− one standard error of the mean

In the REM model, both item and context information are represented as vectors of feature values which are sampled independently from the geometric distribution as follows:

$$P(v) = (1-g)^{v-1}g, \quad v = 1, \ldots, \infty, \tag{1}$$

where g is the parameter for the geometric distribution, v is the actual sampled value for each element in the vector, and the length of the vector was set to 20 for each simulated item.

In the current simulation, only the item noise mechanism of REM was tested. The main reason for that was to investigate the sole effect of item noise on OI and LLE. Of course, adding the context-noise component would further improve the fit of the model to the data. However, the main aim of the current simulation was to evaluate how adding items simultaneously during study and test affects the pattern observed for LLE and OI in item recognition.

In the REM model, the items are stored in memory probabilistically and with some error. The probability of each feature being stored is represented with the parameter u, which determines the strength of encoding. Another parameter, c, moderates the probability of correctly copying the feature value if the feature is stored. If a feature is not correctly copied, then a random feature value is sampled from the same geometric distribution to be stored in the trace.

During retrieval, the test item is compared with all of the traces in memory and a subjective likelihood is calculated using the following equation:

$$\lambda_{(i,j)} = (1-c)^{nq_{(i,j)}} \prod_{v=1}^{\infty} \left[\frac{c + (1-c)g(1-g)^{v-1}}{g(1-g)^{v-1}} \right]^{nm \binom{v,i,j}{}}, \tag{2}$$

where j indexes the test item, i indexes the memory trace, c is the probability of correctly copying the feature value, nq is the number of non-zero feature mismatches, v is the feature value sampled from the geometric distribution, and nm is the number of non-zero feature matches. To make a decision, the subjective likelihood ratios are averaged across traces stored in memory, which results in an odds (Φ) value:

$$\Phi_j = \frac{1}{n} \sum_{i=1}^{n} \lambda_{(i,j)}, \tag{3}$$

where n is the number of traces in the memory search set, i indexes the memory trace, and j indexes the test item. If Φ exceeds the criterion, item j is judged to be "old," if Φ is below the *criterion*, then the item j is judged to be "new." From a signal detection perspective, Φ can be considered as the memory evidence and criterion as the threshold for endorsing an item.

For the no feedback condition, we implemented the encoding during test model as described in Criss et al. (2011). When an item is judged to be old, the best matching episodic trace is updated, which means missing features are likely to being stored. When an item is judged to be new, a new memory trace is added to memory. As a result of these rules, incorrect features are stored in a memory trace for false

alarms and for incorrectly matching targets. Also, sometimes, a second memory trace is stored for missed targets. Therefore, these rules a sharp decrease in HR, while a relatively steady FAR across test blocks.

For the feedback condition, we assumed that when participants received feedback during test, they always added the trace of the test item to memory. Thus, this implementation mimics the list length effect and sometimes second traces are encoded in memory. Because of that, FAR increases and the decrease in HR is less prominent as a function of test block. Table 2 presents the parameter values of the model that was used to simulate data in experiment 2 (see the results of the model in Fig. 3).

General Discussion

The empirical goal of this manuscript was to simultaneously evaluate the effects of adding items during study and adding items during test under identical experimental methodology. We consistently observed a detriment to performance when items were added at any stage and the harm from adding items during study was less than the output interference that resulted from testing items. This pattern of results is consistent with the item noise models in general and specifically with the Criss et al. (2011) implementation of REM. Perhaps the key finding, however, was that the effects of list length and memory testing interacted with the presence of feedback during testing such that feedback disrupts the tendency to erroneously modify existing traces during the commission of false alarm when viewed within the framework extant REM models of recognition memory.

REM predicts a small LLE with higher hits and lower false alarms for shorter lists (e.g., Shiffrin & Stevyers, 1997; Criss and Shiffrin 2004). The Criss et al. (2011) extension incorporates mechanisms during testing as follows. Presenting items at test results in one of two outcomes—remembered items cause updating of the best matching memory trace and non-remembered items cause the storage of a new episodic memory trace. This creates the possibility of three types of additional noise. Miss trials result in the storage of a second copy of a target, false alarm trials result in a memory trace being updated with incorrect information (i.e., features that belong to a different item), and some hit trials may also result in item features being stored in the wrong trace (i.e., when the best matching trace does not correspond to the target being tested but to a different target item). Presumably, the latter type of error is minimal, but nonetheless, it is possible. These mechanisms of encoding during retrieval respects the principles of REM that encoding is noisy and error-prone and is consistent with the larger literature showing that memories are malleable and sometimes memories of different events are combined. These errors all contribute to a decrease in the HR across test trial. The pattern of FAR is variable because it depends on the relative degree of each storage mechanism. Storing additional traces (as in misses and correct rejections) increases the FAR, whereas updating traces (as in hits and false alarms) decreases the FAR.

Table 2 Parameter values used in the REM model simulations

Parameter	Criss et al. (2011)	Experiment 2
Number of features	20	20
g	0.35	0.35
c	0.7	0.7
u	0.16	0.22
Criterion	0.72	0.75
Study list length	75	
Long study list length		200
Short study list length		75
Test list length	150	150

In addition to causing the particular pattern of OI that is observed in human behavior, updating traces also results in differentiation-based findings of a null list strength effect and the strength-based mirror effect. Both of these findings reflect the fact that strengthening memories reduces the noise and subsequent confusion associated with those memories. This connection between otherwise unrelated empirical regularities suggests that differentiation is a core mechanism underlying memory (see also Koop & Criss, 2015; Kilic et al., 2017).

Updating remembered traces serves as a potential mechanism for reconsolidation. Memories that are reactivated are brought into a state where they can be modified or updated, and then, the new changed memory is reconsolidated (Barry and McGuire 2019; Elsey et al. 2018; Hardt et al. 2010). REM offers a mechanism for this phenomenon and offers predictions. For example, memories that are well encoded are more likely to be retrieved but have little less to gain from reconsolidation, whereas memoires that are minimally encoded have much to gain from reactivation and reconsolidation (e.g., see Kılıç et al. 2017). A fruitful line of future research could be to explore the relationship between REM, these hallmark findings in recognition memory, and reconsolidation.

Feedback presented during test served as a moderator. When feedback was presented at test, the OI in HR was smaller and the size of the LLE was larger compared with when feedback was withheld. To ensure that the apparent effects of feedback were not due to storing any item or due to simple sensory input, we showed the word "next" in place of the feedback in experiments 2 and 3 and the pattern held. The role of feedback was novel because a number of studies show that providing

feedback does not change accuracy and the same was true here. Overall accuracy was unaffected by feedback. However, feedback changed the pattern of hit rates and false alarm rates as they relate to the pattern of OI and LLE. Within the framework of the REM model, this suggests that providing feedback causes new traces to be added to memory rather than, or in addition to, updating traces during test. One possibility for this pattern is that whenever feedback indicates that an 'old' response was incorrect a new trace is stored in response to prediction error. The worse performance, the more incorrect responses and the more additional traces are added. Long lists have more such errors resulting in more traces added during test than short lists. Another possibility is that updating proceeds just as described by Criss et al. (2011). In addition to that, participants store a trace including the feedback itself. In either of these cases, additional learning occurs during test rather than occurs during study, consistent with the testing effect (Karpicke and Roediger 2008; Karpicke et al. 2014) and the benefits of testing are potentially more pronounced when feedback is provided (cf, Aue, Criss, & Prince, 2015).

This mechanism of feedback, adding new traces, resulted in an increase in FAR and dampened the decrease in HR because especially in HR a second copy of a trace increased the overall memory noise by activating common random features. Therefore, adding a new trace after a feedback caused multiple traces (two in this case) of the same study item. However, in the current study, the effect of feedback on a second test of the same item was not measured, which might be an interesting question for the role of feedback on item recognition in future studies. Building on that, the role of feedback in multiple repetitions can be further investigated to understand the underlying mechanisms of the spacing effect such as recursive reminding or the modification of the initial trace (e.g. Benjamin and Tullis 2010; Hintzman 2010; Wahlheim et al. 2014).

A third possible explanation has nothing to do with adding information to episodic memory. Perhaps feedback serves to better isolate the relevant subset of memory to search. If isolating the relevant context is imperfect and extra-list memory traces enter the set, then any effect of the list itself is diluted. Better isolating the study list amplifies manipulations of the study list such as the length of the list. Evidence for this idea comes from the fact that false alarms seem to be more affected by feedback than hits, mainly due to false alarms being particularly sensitive to extra-list traces because the foils could be included in that set.

Alternatively, if performance decreases as a function of test position because of the drifting context over the course of testing, then it is possible that receiving feedback increases the rate of contextual drift. A recently developed model (Osth et al. 2018) aims to examine how much noise in recognition testing comes from the items presented in the experiment and how much noise from the drifting context over the course of the experiment. Osth et al. (2018) suggested that the main cause of decreasing performance is in fact drifting context while other items in memory have minimal effect on subsequent testing. The role of feedback in this study can also be explained by an increase in the rate of drifting context. As simulated by Osth et al. (2018), drift in context results in an increase in FAR as observed in the feedback conditions; however, it also results in a decrease in HR, which was not the case in the current study. Because the model proposed by Osth et al. (2015) is a dynamic model which benefits from conjoint measures of reaction time and response rates, including a change in speed accuracy parameters might suggest a better understanding on the role of feedback.

Discriminating between these and other possibilities is an avenue left for future research that is focused on the role of feedback in memory. For example, testing memory for the feedback provided to any item might be informative. In addition, analyzing memory for the tested items conditional on response given for the initial test might help to discriminate between these possibilities.

Data Availability Data is available at https://osf.io/8zzbm/.

References

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38*(4), 341–380.

Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language, 85*, 27–41. https://doi.org/10.1016/j.jml.2015.06.001.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Barry, D. N., & McGuire, E. A. (2019). Remote memory and the hippocampus: a constructive critique. *Trends in Cognitive Science, 23*(2), 128–142.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228–247.

Brandt, M., Zaiser, A.-K., & Schnuerch, M. (2019). Homogeneity of item material boosts the list length effect in recognition memory: a global matching perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(5), 834–850. https://doi.org/10.1037/xlm0000594.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*(2), 231–248.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: similarity and the strength based mirror effect. *Journal of Memory and Language, 55*(4), 461–478.

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59*(4), 297–319.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 484.

Criss, A. H., Salomão, C., Malmberg, K. J., Aue, W. R., Kılıç, A., & Claridge, M. (2017). Release from output interference in recognition memory: a test for the attention hypothesis. *The Quarterly Journal of Experimental Psychology, 71*, 1081–1089. https://doi.org/10.1080/17470218.2017.1310265.

Criss, A. H., & Koop, G. J. (2015). Differentiation in episodic memory. Cognitive modeling in perception and memory: a Festschrift for Richard M. Shiffrin, 112–115.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64*(4), 316–326. https://doi.org/10.1016/j.jml.2011.02.003.

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: a comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55*(4), 447–460.

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: a comment on Dennis and Humphreys (2001). *Psychological Review, 111*(3), 800–807. https://doi.org/10.1037/0033-295X.111.3.800.

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

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–478.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: the case of the list-length effect. *Journal of Memory and Language, 59*(3), 361–376. https://doi.org/10.1016/j.jml.2008.06.007.

Elsey, J. W. B., van Ast, V. A., & Kindt, M. (2018). Human memory reconsolidation: a guiding framework and critical review of the evidence. *Psychological Bulletin, 144*(8), 797–848.

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language, 110*, 104065. https://doi.org/10.1016/j.jml.2019.104065.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1355–1369. https://doi.org/10.1037/0278-7393.20.6.1355.

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: evidence for adaptive criterion learning. *Memory & Cognition, 36*(4), 703–715.

Hardt, O., Einarsson, E. Ö., & Nader, M. (2010). A bridge over troubled water: reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annual Review of Psychology, 61*, 141–167.

Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition, 38*(1), 102–115.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38*(4), 389–406.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: an episodic context account. *Psychology of Learning and Motivation, 61*, 237–284.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

Kılıç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology, 92*, 65–86. https://doi.org/10.1016/j.cogpsych.2016.11.005.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39*(2), 348–363.

Koop, G. J., Criss, A. H., & Malmberg, K. J. (2015). The role of mnemonic processes in pure-target and pure-foil recognition memory. *Psychonomic Bulletin & Review, 22*(2), 509–516.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23*(2), 115–119. https://doi.org/10.1177/0956797611430692.

Malmberg, K. J., Holden, J. E., & Shiffren, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 319.

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory, *Journal of Experimental Psychology: Learning, Memory, and Cognition., 28*(4), 616.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724–760.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*(6), 609–626.

Murdock, B. B. & Anderson, R. E. (1975). Encoding, storage, and retrieval of item information. In Solso Robert L (Ed.), Information processing and cognition: the Loyola Symposium. Lawrence Erlbaum.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 384.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260.

Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology, 104*, 106–142.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 163.

Ratcliff, R., & Hockley, W. E. (1980). Repeated negatives in item recognition: nonmonotonic lag functions. *Attention and performance VIII*. Hillsdale*:* Erlbaum.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: a computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 294. https://doi.org/10.1037/0278-7393.26.2.294.

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning and Memory, 6*(1), 91.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166.

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-

based mirror effect in recognition memory. *Journal of Memory and Language, 63*(1), 18–34.

Strong Jr., E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review, 19*(6), 447–462.

Swets, J. A., & Green, D. M. (1961). Sequential observations by human observers of signals in noise. The University Press.

Wahlheim, C. N., Maddox, G. B., & Jacoby, L. L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: sufficient but not necessary. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 94.

Wickens, D. D. (1970). Encoding categories of words: an empirical approach to meaning. *Psychological Review, 77*(1), 1–15. https://doi.org/10.1037/h0028569.