Memory

ISSN: 0965-8211 (Print) 1464-0686 (Online) Journal homepage: http://www.tandfonline.com/loi/pmem20

Filling in the gaps: using testing and restudy to promote associative learning

John B. Bulevich, Ayanna K. Thomas & Charles Parsow

To cite this article: John B. Bulevich, Ayanna K. Thomas & Charles Parsow (2016) Filling in the gaps: using testing and restudy to promote associative learning, Memory, 24:9, 1267-1277, DOI: 10.1080/09658211.2015.1098706

To link to this article: https://doi.org/10.1080/09658211.2015.1098706

MEMORY, 2016
VOL. 24, NO. 9, 1267–1277
http://dx.doi.org/10.1080/09658211.2015.1098706

Routledge
Taylor & Francis Group

Filling in the gaps: using testing and restudy to promote associative learning

John B. Bulevich[a], Ayanna K. Thomas[b] and Charles Parsow[b]

[a]Department of Psychology, Stockton University, Galloway, NJ, USA; [b]Department of Psychology, Tufts University, Medford, MA, USA

ABSTRACT

Although testing has been shown to potentiate subsequent learning [Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, 18, 879–919.], the mechanisms that influence this effect are not entirely understood. The present research examined the relationship between associative binding and test-potentiation effects. We hypothesised that test-potentiation effects would be most pronounced when participants could easily extract the relationship among word groupings. Towards that end, we compared three-word groupings, or triads, that were either semantically related or unrelated. Participants engaged in repeated study, repeated testing, or engaged in interpolated study and test prior to a final test. Final test performance was greatest for participants who engaged in interpolated study and test on related triads. The results support three primary conclusions: (1) testing aids in associative binding; (2) associative binding is facilitated by retrieval practice and restudy pairings; and (3) pre-existing associations facilitate test-potentiation effects.

ARTICLE HISTORY
Received 16 February 2015
Accepted 11 September 2015

KEYWORDS
Retrieval; testing

Research has consistently demonstrated that repeated testing improves long-term retention as compared to repeated study (for review see Roediger & Butler, 2013; Roediger & Karpicke, 2006a). The beneficial effects of repeated testing have been demonstrated in studies using word lists (Zaromb & Roediger, 2010), paired associates (Toppino & Cohen, 2009), and even complex, educationally relevant material (Butler, 2010; Roediger & Karpicke, 2006b; McDaniel, Howard, & Einstein, 2009). All of these examples demonstrate situations where testing leads to better long-term retention as compared to solely restudying. However, in most situations outside of the laboratory, testing occurs in tandem with re-exposure to the previously studied and new related material. That is, testing rarely occurs in isolation of study. The present study is primary interested in how testing influences subsequent learning.

The concept of testing enhancing learning is not new (Izawa, 1966). Izawa used lists of consonant-vowel-consonant trigrams (ZAK) and paired them with either two digit numbers (40) or high frequency nouns (NORTH). These pairs were repeatedly studied or interspersed with tests. The results were explained in the context of reinforcement models (Estes, 1955). When the lists were re-presented after a test, participants' performance was better on subsequent tests compared to conditions where interspersed testing had not occurred. Similar results emerged examining interference effects using AB-AD word pairs. When AB pairs were tested prior to the presentation of AD pairs, participants learned the AD pairs better (as measured by a modified-modified free recall test) compared to when participants had not been given that initial test (Tulving & Watkins, 1974).

In more recent research, this general principle of test potentiated learning has been applied to a number of different situations. Using more complex materials, research has demonstrated that retrieval followed by restudy more than doubled memory performance compared to study alone (Karpicke & Roediger, 2010). Similarly, when people took tests and then had an opportunity to restudy related material, performance was better than when they did not take those initial tests (Butler, 2010). Recently, in the context of the misinformation effect, testing of the original event produced improved memory for the misleading post event information (Gordon, Bulevich, & Thomas, 2015; Gordon & Thomas, 2014).

The aforementioned research has demonstrated that testing enhances retention and potentiates new learning. In the present research, we tested the relationship between test potentiation and associative binding. We hypothesised that interpolated study and test would facilitate associative binding through a potentiation mechanism. In this context, we define associative binding as the ability to form and retrieve links among single units of information (Naveh-Benjamin, 2000). Our hypothesis emerged from research that suggests that retrieval and even failed retrieval may potentiate learning through a feedback mechanism. That is, testing may highlight information that

CONTACT John B. Bulevich john.bulevich@stockton.edu

requires additional study (Arnold & McDermott, 2013; Gordon et al., 2015; Izawa, 1970, 1971; Thompson, Wegner, & Bartling, 1978). Retrieval may also enhance the recollection of source-specifying details (Chan & McDermott, 2007; Szpunar, McDermott, & Roediger, 2008) Research has also demonstrated that testing may serve to contextually separate the two learning episodes thereby reducing interference between learning episodes and fostering integration within episodes (Szpunar et al., 2008; for similar ideas see O'Reilly & Rudy, 2000). All of these findings point to the conclusion that testing provides useful information regarding what has been learned from a previous episode, and that information may improve learning when given the opportunity for restudy.

Thus, retrieval may facilitate learning and memory in several ways. The act of retrieval may serve as an additional learning or encoding episode (Roediger & Nestojko, 2015). The act of retrieval may highlight gaps in knowledge (Izawa, 1966). Finally, retrieval practice likely strengthens the association between individual items (associative binding), and facilitates the linking of those items to pre-existing knowledge. Research suggests that new information is more easily learned if individuals have an established knowledge structure in which to integrate that information (e.g., Kintsch & Greene, 1978; Kintsch & Kintsch, 1978). For example, when participants studied narratives that had a familiar structure, their memory for those narratives were better than for ones with atypical structures. Kintch and Greene argued that the familiar structure allowed for better integration which resulted in better memory. Similarly, Chan (2009) showed that when participants were encouraged to make connections between different facts within a prose passage, repeated testing had a more pronounced effect on final retention than when participants were encouraged to try and memorise individual facts. In the present study, we hypothesise that testing should promote associative binding, and that new related concepts are more easily incorporated into an already established schema or knowledge structure.

Associative binding should be facilitated by restudy following testing. That is, testing may highlight weak binding, and restudy may allow participants to strategically regulate learning to promote stronger associations between individual units. Thus, these effects are dependent on the relationship between testing and restudy. However, while testing should facilitate binding overall, materials that have a priori associations should benefit more from testing and restudy than materials that do not. These predictions build on recent work examining test-potentiation effects using pre-testing procedures (testing prior to any study) that demonstrated enhanced final test performance for strongly related items relative to weakly related items (Grimaldi & Karpicke, 2012; Yan, Yu, Garcia, & Bjork, 2014).

The present study had two primary motivations. First, the series of experiments examined how retrieval practice promotes associative binding. Across experiments, we varied test format in order to determine whether retrieval practice facilitated flexible access to studied material, and to determine whether retrieval promoted recollection-based responding. Second, we tested the hypothesis that improved final test performance is due to the associative binding effects which resulted from initial retrieval. In the context of the current set of experiments, participants studied a series of triads (DIAMOND-RUBY-EMERALD). Each word of the triad served as the single unit. Successful performance on subsequent memory tests required that participants not only retrieve individual units of triads, but also form and retrieve associative links among members of triads. In a triad (A-B-C) multiple associations must be formed in contrast to paired associates (as has primarily been explored in the previously cited recent work; Grimaldi & Karpicke, 2012; Karpicke & Roediger, 2008; Yan et al., 2014) with only a single associative link. Participants may bind some or all of the elements of the triad into a single unit. Elements that are not bound should be identified during an initial test, and when given the opportunity to restudy, participants may be better able to direct attention to those elements, facilitating learning (cf., Arnold & McDermott, 2013; Gordon et al., 2015). We employed triads to increase the chances of partial retrieval failure after initial study. We hypothesised that gaps in knowledge that remained after the first study episode would be identified in subsequent testing phases. Thus, retrieval failure would serve as a form of feedback and should facilitate effective deployment of attention to unlearned material upon restudy. In addition, materials that have pre-existing semantic associations will benefit more from interpolated testing and study as compared to unassociated information, because targeted restudy will be able to capitalise on prior knowledge. For example, if one studies a grouping of words that are related to the concept of gemstones and remembers the general association but has not learned all of the individual items, restudy can be directed at filling in specific gaps and associating individual items not only to each other, but to the superordinate concept. Finally, we also predicted that in cases where study alone is compared with testing alone, testing should lead to improved final test performance.

Experiment 1

Participants

Seventy-two undergraduate students from Tufts University participated in this experiment for course research credit. Participants were equally distributed across the three groups.

Materials and procedure

We employed a 3 (Practice: Study-Study-Study, Study-Test-Study, Study-Test-Filler) X 2 (Triad Type: Related, Unrelated) mixed variable design. Practice was manipulated between

participants and triad type was manipulated within participants. Participants first studied a list of 24 triads, or group of three words. Twelve triads consisted of semantically related words, and 12 consisted of semantically unrelated words. Unrelated triads consisted of words that were completely unrelated to one another, as measured by low forward and backward associative strength. Semantically related triads consisted of words that were all related as measured by high forward and backward associative strength and that were highly typical members of the same semantic category as determined by the Van Overshelde, Rawson, and Dunlosky (2004) category norms. Associative strength was determined using the University of South Florida Free Association Norms list (Nelson, McEvoy, & Schreiber, 1998). For related triads, each of the six pairs within a triad had an average forward associative strength of 0.55 and an average backward associative strength of 0.48. For unrelated triads, relationships between pairs in a given triad were lower than 0.03.

For the Study-Study-Study group (SSS), a given triad unit was selected at random and presented collinearly in the middle of the computer screen for five seconds. A five minute retention interval followed the first study phase. During this time, participants played a game of Sudoku. Participants were represented with the identical list of triads units for restudy followed by a second five-minute retention interval, and a third study phase. In each of the study phases, participants were presented with the same triad units; however order of presentation of units was randomised. Thus, while triad units were identical among study sessions, order of presentation varied among study sessions and participants. At the end of this first session, participants were reminded to return 48 h later for final testing. After the 48-h delay, participants took a forced cued recall test in which the first word of previously studied triads was presented and participants were instructed to provide the two associated words.

Initial study was identical to the SSS group for participants in the Study-Test-Study group (STS). After study, participants in the STS group took an immediate forced cued recall test after the first retention interval. Cued recall testing consisted of providing the first word of a triad with instructions to produce the two associated targets. Test cues were presented in a random order. After this first test, participants in the STS group were presented with all of the triads for restudy.

For the Study-Test-Filler (STF) group, participants engaged in the identical initial study and testing procedure as described for the STS group. After the first cued recall test, participants in the STF group, played Sudoku for five minutes instead of restudying. All groups received the same final forced cued recall test after a 48 h delay. Thus, the final session was identical across the three experimental groups. All cued recall tests (Initial, Final) were identically constructed and self-paced. That is, participants were presented with the first words of triads as cues, and were instructed to recall the two associated targets.

Results

Performance on the Initial Test and Final Test were scored using two criteria. When the strict scoring criterion was used, a response was counted as correct if both targets were correctly produced in association with the appropriate cue. When the liberal scoring criterion was used, a response was counted as correct if one target was correctly produced in association was a cue. Average Initial test performance was subjected to a 2 (Practice: STS, STF) × 2 (Triad Type: Related, Unrelated) analysis of variance (ANOVA). Two separate ANOVAs were conducted using calculated accuracy under the strict and liberal scoring criteria, respectively. Performance was statistically identical on the Initial Test for participants in the STS and STF groups regardless of scoring criteria, F's < 1. However, for both scoring criteria, main effects of triad type were found, [liberal: $F(1, 46) = 105.96$, $p < .001$, $\eta_p^2 = .69$; strict, $F(1, 46) = 47.98$, $p < .001$, $\eta_p^2 = .51$]. The interaction between the two variables was not significant under either scoring criteria. Table 1 illustrates that participants remembered more related than unrelated items on the Initial Test.

Turing now to performance on the Final Test, a 3 (Practice: SSS, STS, STF) x 2 (Triad Type: Related, Unrelated) mixed design ANOVA using the liberal scoring criteria found a main effect of Triad Type, $F(1, 69) = 120.61$, $p < .001$, $\eta_p^2 = .63$, and a main effect of Practice, $F(2, 69) = 3.81$, $p < .05$, $\eta_p^2 = .10$. As Table 1 demonstrates, participants were more accurate when recalling related ($M = .28$) as compared to unrelated ($M = .11$) triads. In addition, participants in the STS group performed better ($M = .32$) than participants in both the STF ($M = .22$) and SSS ($M = .19$) groups. Finally, the interaction between Triad Type and Practice was significant, $F(2, 69) = 3.70$, $p < .05$, $\eta_p^2 = .10$.

Separate one-way ANOVAs for each Triad Type were performed to deconstruct this interaction. Importantly, the difference between practice groups was not significant when unrelated triad groups were analysed, $F < 1$. For unrelated triad groups, participants demonstrated low, but above floor performance ($M = .11$). For related triads, group differences emerged, $F(2, 69) = 4.76$, $p < .05$, $\eta_p^2 = .12$. After a Bonferroni correction, pairwise comparisons revealed that participants in the STS group were more accurate than those in the SSS, $t(46) = 2.77$, $d = .81$, and those in the STF groups, $t(46) = 1.98$, $d = .62$.

A similar pattern of results was demonstrated under the strict scoring criteria. Specifically, a 3 (Practice: SSS, STS, STF) × 2 (Triad Type: Related, Unrelated) mixed design ANOVA found a main effect of Triad Type, $F(1, 69) = 38.85$, $p < .001$, $\eta_p^2 = .36$, and a main effect of Practice, $F(2, 69) = 4.92$, $p < .01$, $\eta_p^2 = .13$. Finally, the interaction between Triad Type and Practice was significant, $F(2, 69) = 7.28$, $p < .001$, $\eta_p^2 = .17$. Separate one-way ANOVAs for each Triad Type were performed to deconstruct this interaction. As with the liberal criteria, the difference between groups was not significant for when unrelated triad

Table 1. Experiment 1—Average cued recall performance for related and unrelated triads under two scoring criteria.

	SSS	STF	STS
Initial test			
Liberal			
Related		.42 (.04)	.51 (.05)
Unrelated		.14 (.03)	.15 (.04)
Strict			
Related		.22 (.03)	.20 (.04)
Unrelated		.06 (.02)	.04 (.02)
Final test			
Liberal			
Related	.29 (.05)	.36 (.04)	.51 (.06)
Unrelated	.09 (.03)	.09 (.01)	.14 (.04)
Strict			
Related	.08 (.02)	.14 (.03)	.28 (.05)
Unrelated	.05 (.02)	.03 (.01)	.08 (.03)

Note: Standard errors are in parentheses.

groups were analysed, $F = 1.0$. For related triads, group differences emerged, $F(2, 69) = 6.83$, $p < .05$, $\eta_p^2 = .17$. After a Bonferroni correction, pairwise comparisons revealed that participants in the STS group were more accurate than those in the SSS, $t(46) = 3,39$, $d = 1.00$, and those in the STF groups, $t(46) = 2.24$, $d = .63$.

Finally, we performed an analysis examining performance on the Final Test in reference to performance on the Initial Test. The goal of this contingency analysis was to examine gains and losses between the Initial Test and the Final Test. This analysis was employed to provide stronger evidence that testing followed by restudy promoted new learning. Net gains were calculated by adding the number of new targets produced on the Final Test. Net losses were calculated by adding the number of targets reported in the Initial Test that were not re-reported on the Final Test. We compared net gains using a 2 (Triad Type: Related, Unrelated) x 2 (Practice: STS, STF) mixed design ANOVA. A main effect of Triad Type was found, $F(1, 46) = 17.51$, $p < .001$, $\eta_p^2 = .309$. A main effect of Practice was found, $F(1, 46) = 14.07$, $p < .001$, $\eta_p^2 = .23$. As Figure 1 illustrates, participants in the STS group demonstrated greater gains than participants in the STF group. When losses were analysed we again found a main effect of Triad Type, $F(1, 46) = 31.89$, $p < .001$, $\eta_p^2 = .41$. There was no effect of Practice, nor did Practice and Triad Type interact when losses were analysed.

Experiment 1 discussion

The most important finding from Experiment 1 was that when study and test were interpolated, performance was better than with either repeated studying or retrieval practice with no opportunity to restudy following initial encoding. This was true under both strict and liberal scoring criteria. These results also demonstrated that testing followed by restudy resulted in greater net gains between Initial and Final testing as compared to testing in the absence of restudy. This finding suggests that testing followed by restudy facilitated learning but did not reduce forgetting. This finding is consistent with previous research that suggests that testing may serve to highlight gaps in knowledge, resulting in strategic encoding of previously unlearned information when given the opportunity to restudy (see Arnold & McDermott, 2013; Chan, Wilford, & Hughes, 2012; Gordon et al., 2015; Gordon & Thomas, 2014; Izawa, 1970, 1971; Thomas, Bulevich, & Chan, 2010; Thompson et al., 1978). In addition, pre-existing associations facilitated test-potentiation effects. Specifically, for unrelated triads, the schedule of study and test did not affect Final test performance, whereas for related triads, Final test performance was greatest when participants engaged in interpolated study and testing. These findings suggest that pre-existing associations may facilitate test-potentiated learning. In the present study, taking the Initial test may have clearly delineated what information was and was not encoded. Restudy could then be targeted to the less well encoded items; however learning may have been more likely to occur when participants could easily associate items both to one another and to a superordinate concept. These ideas are also similar to those expounded in the pre-testing education literature. That is, research has demonstrated that providing students with a series of pre-study questions facilitated learning (Hamaker, 1986; Lewis & Mensink, 2012).

Experiment 2

Given the results of the first experiment, Experiment 2 had two primary goals. The first was to examine whether the format of the test was a critical factor for the potentiating of new learning. Previous research has demonstrated that tests with greater retrieval demands (recall; production based tests) resulted in larger testing effects compared to tests with fewer retrieval demands (recognition) (Halamish & Bjork, 2011). Experiment 2 investigated whether interpolated study and test would result in better overall retention on a final test when the retrieval demands of testing were reduced. In other words, we wanted to determine whether the reduced demands of the recognition test would still lead to the potentiation effects observed in Experiment 1. The second was to directly test the development of associations across triad groupings by using an associative recognition test. Associative recognition tests, where participants study pairs and must be able to discriminate intact from recombined pairs, have been utilised to examine the issue of associative vs. item memory (see Naveh-Benjamin, 2000). The associative recognition test used in Experiment 2 was designed to assess associative binding in addition to individual item memory. While one may use familiarity or some form of retrieval fluency to correctly recognise an individual item, to correctly recognise a pair or triad, the unit must have been bound or integrated. We hypothesised that interpolated testing and study would facilitate better binding of the entire triad unit as compared to repeated study or repeated testing alone.

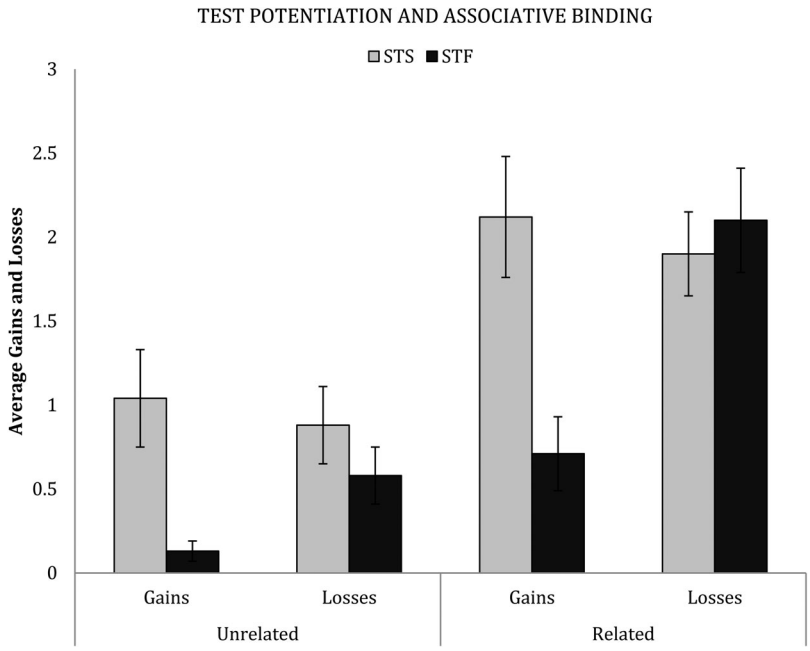

Figure 1. Average gains and losses (items within a triad) between the Initial and Final test for experiment 1.

In the present experiment, we examined false alarms associated with different categories of lures. The recognition test was such that participants were presented with the first word of the triad as a cue and were required to correctly recognise the two associated targets. Lures varied by manipulating one or both of the targets. A target that had been previously studied with the cue could be presented with a new target. Alternatively, two targets that had been previously studied in a different context could be presented with a new cue at test. Thus, all lures consisted of previously studied cues and targets. We predicted that testing followed by restudy would facilitate performance (primarily through a reduction of false alarms) on this associative recognition test.

Method

Participants

Eight-four undergraduate students from Tufts University participated in this experiment for course research credit. Participants were equally distributed across the three testing groups.

Materials and procedure

The experimental design and procedure was identical to Experiment 1, with two exceptions. First, participants studied 20 triad groups in Experiment 2 as opposed to 24. We reduced the number of total triads in order to improve memory for unrelated triads. Second, participants were given a recognition test as opposed to a cued recall test. Participants studied groups of three related or unrelated words. After initial study, two groups of participants received a recognition test while a third group engaged in repeated study.

At test, participants were required to make a yes/no decision for 80 triad groups. Twenty triads had been previously studied and 60 served as different classes of lures. There were four lure classes. Each lure was designed such that a studied cue 'A' was presented with a new combination of 'B' and 'C' lures. Some items that served in lure trials had been previously studied. The lure groups were as follows: cue with old target and new target (ON), cue with new target and old target (NO), cue with two new targets that had not been presented together ($NN_{unpaired}$), and cue with two new targets in which the two targets had been presented together (NN_{paired}).

Lures were generated independently for related and unrelated triads. Specifically, a cue for a related triad would not be paired with lures from unrelated triads as well as the converse. The 80 triad groups were randomly presented at test. Participants pressed the 'Y' key if they recognised a triad and the 'N' key if they did not recognise a triad. Participants were told to respond 'Y' only if the grouping of three words was exactly what they remembered as having studied. After taking the first test, participants in the STS group were given the opportunity to restudy. After taking the first test, participants in the STF group were given a Sudoku puzzle to complete. After the first restudy, participants in the SSS group were given a third study of the triads. Between each phase, participants engaged in a five-minute filler task in which they played Sudoku. After a 48 h retention interval, participants returned to the lab for the second session, which consisted of one test.

Experiment 2 results

Test 1

On the Initial Test, we first examined hit rates in the context of a 2 (Triad Type: Related, Unrelated) × 2 (Practice: STS, STF) mixed design ANOVA. Only the main effect of Triad Type was significant, $F(1, 54) = 108.72$, $\eta_p^2 = .69$. The proportion of studied triads recognised was higher for related triads ($M = .82$) as compared to unrelated triads ($M = .51$). False alarms were analysed using a 2 (Triad Type: Related, Unrelated) × 2 (Practice: STS, STF) × 3 (Lure Type: NN_{paired}, $NN_{unpaired,}$ NO) mixed design ANOVA. We collapsed across the new-old/old-new lure types, because a preliminary analysis revealed no difference in false alarm responding based on this lure classification. We found a main effect of Lure Type, $F(12, 108) = 17.44$, $p < .001$, $\eta_p^2 = .24$. False alarms were higher for NO ($M = .13$) and NN_{paired} ($M = .13$) as compared to $NN_{unpaired}$ ($M = .08$) lures. The interaction between Triad Type and Lure Type was also significant, $F(2, 108) = 5.22$, $p < .01$, $\eta_p^2 = .10$. Separate one-way ANOVAs for each Triad Type were conducted to decompose this interaction. As demonstrated in Table 2, on the Initial Test, lures did not significantly differ for related triads. However, for unrelated triads participants were more likely to false alarm to NN_{paired} ($M = .21$) as compared to $NN_{unpaired}$, ($M = .12$) lures, $t(55) = 5.23$, $d = .62$. Similarly, participants were more likely to false alarm to NO ($M = .19$) lure pairings as opposed to $NN_{unpaired}$, lures, $t(55) = 4.08$, $d = .56$. The difference between false alarms to NO and $NN_{unpaired}$ lures did not reach statistical significance, $t < 1$. All pairwise comparisons were made using a Bonferroni correction. Thus, when participants were presented with lures that had been previously studied together, they were more likely to false alarm than when presented with previously unpaired lures.

Final test

As with the Initial Test, we examined hits and false alarms separately for the Final Test. A 2 (Triad Type: Related, Unrelated) × 3 (Practice: SSS, STS, STF) mixed designed ANOVA on hit rates yielded a main effect of Triad Type, $F(1, 81) = 127.43$, $p < .001$, $\eta_p^2 = .61$. As Table 2 demonstrates, participants correctly recognised more related than unrelated triads. A main effect of group was also found, $F(2, 81) = 16.29$, $p < .001$, $\eta_p^2 = .29$. Participants in the STS group correctly recognised more triads than participants in the STF groups, $t(54) = 5.18$, $d = 1.46$. Participants in the SSS group also recognised more triads correctly than participants in the STF group, $t(54) = 4.03$, $d = 1.07$. With the Bonferroni correction, the difference between the SSS and STS groups was not significant.

Table 2. Experiment 2—Proportion of hits and false alarms by practice group.

	Hit	New-New (P)	Old-New	New-New (U)
Initial test				
Related				
STS	.86 (.13)	.04 (.06)	.06 (.08)	.02 (.01)
STF	.76 (.19)	.04 (.15)	.08 (.10)	.05 (.02)
Unrelated				
STS	.50 (.04)	.19 (.03)	.16 (.02)	.09 (.02)
STF	.52 (.03)	.23 (.03)	.23 (.02)	.15 (.02)
Final test				
Related				
SSS	.88 (.02)	.13 (.02)	.16 (.02)	.14 (.02)
STS	.91 (.02)	.01 (.01)	.02 (.01)	.01 (.01)
STF	.74 (.04)	.08 (.03)	.08 (.02)	.02 (.01)
Unrelated				
SSS	.50 (.04)	.30 (.03)	.34 (.03)	.13 (.02)
STS	.63 (.05)	.12 (.02)	.20 (.02)	.08 (.01)
STF	.37 (.04)	.20 (.04)	.20 (.02)	.17 (.04)

Note: Standard error is in parentheses.

While important, the examination of hits is less relevant to the primary hypotheses that motivated the present study. In Experiment 2 lures were constructed to test the hypothesis that interpolated study and testing would facilitate stronger associative binding across triad units as compared to isolated study or testing. Specifically, we expected to see low levels of false alarms across all lure types for participants who engaged in interpolated study and test. However, in cases where participants engaged in isolated study or isolated testing, we expected higher false alarms to lures that consisted of two previously paired words. To examine test-potentiated binding we compared the three lure types in a 2 (Triad Type: Related, Unrelated) × 3 (Practice: SSS, STS, STF) × 3 (Lure Type: NN_{paired}, $NN_{unpaired,}$ NO) mixed design ANOVA. Main effects for Practice, $F(2, 81) = 16.01$, $p < .001$, $\eta_p^2 = .28$, Triad Type, $F(1, 81) = 108.81$, $p < .001$, $\eta_p^2 = .57$, and Lure Type, $F(2, 162) = 22.28$, $p < .001$, $\eta_p^2 = .22$, were found. However, these main effects were qualified by the three-way interaction found among tested variables, $F(4, 162) = 5.82$, $p < .001$, $\eta_p^2 = .13$. To decompose this interaction two separate 3 (Practice: SSS, STS, STF) x 3 (Lure Type: NN_{paired}, $NN_{unpaired,}$ NO) mixed design ANOVAs were conducted for each Triad Type. For related triads, only a main effect of Practice was found, $F(2, 81) = 20.23$, $p < .001$, $\eta_p^2 = .33$. Participants in the SSS ($M = .14$) group were significantly more likely to false alarm than participants in the STS ($M = .01$) and STF ($M = .05$) groups. Pairwise comparisons using a Bonferroni correction demonstrated that false alarm rates were lower in the STS group as compared to the STF group, $t(54) = 2.03$ $d = .60$. False alarm rates in the STS group were lower than those in the SSS group, $t(54) = 4.45$, $d = 1.29$, False alarm rates did not differ between the SSS and STF groups, *ns*. Interestingly, for unrelated lures the pattern of false alarms was quite different. As with related triads, we found a main effect of Practice, $F(2, 81) = 7.95$, $p < .001$, $\eta_p^2 = .16$. However, we also found a main effect of Lure Type, $F(2, 162) = 21.55$, $p < .001$, $\eta_p^2 = .21$. In addition, the interaction between the two variables was significant, $F(4, 162) = 5.50$, $p < .001$, $\eta_p^2 = .12$, The pattern of false alarms was affected by group (Practice) activity. Specifically, participants in the STF group demonstrated statistically identically false alarms across all lure types ($M = .19$). In contrast, participants in the SSS group demonstrated lower false alarm rates for $NN_{unpaired}$, lures ($M = .13$)

as compared to NN_{paired} ($M = .31$) and NO lures ($M = .35$), [$NN_{unpaired}$ – NN_{paired}: $t(27) = 6.55$, $d = .96$; $NN_{unpaired}$ – NO: $t(27) = 6.51$, $d = 1.20$]. The difference between NN_{paired} and NO lures was not significant. Finally, participants in the STS group demonstrated higher false alarms for NO lures ($M = .20$) as compared to both $NN_{unpaired}$ ($M = .08$) or NN_{paired} ($M = .12$) lures, [NO – $NN_{unpaired}$: $t(27) = 4.97$, $d = 1.14$; NO – NN_{paired}: $t(27) = 2.61$, $d = .68$]. The difference between $NN_{unpaired}$ and NN_{paired} lures was not significant. Comparisons were significant after Bonferroni corrections.

Experiment 2 discussion

The results of Experiment 2 supported the hypothesis that interpolated study and test facilitated the associative binding of the triads. Strong associative binding was demonstrated by high hit rates and low false alarm rates. Interestedly, only the STS group demonstrated high hits and low false alarms for all lure types on the Final test. This was found for both triad types. Contrast this pattern with that observed in the SSS group. Repeated study did positively impact recognition hits; however, false alarms across all lures types was high for this group. Retrieval practice, whether interpolated or isolated, reduced false alarm responding. These results suggest that taking an initial test facilitates binding of cues with targets, consistent with previous literature (Carpenter, Pashler, & Vul, 2006) and promotes recollection-based responding. These results also suggest that interpolated testing and study resulted in the strongest binding across all three elements of a triad for both related and unrelated groups, and likely facilitated better source-monitoring and enhanced recollective processes (i.e., Chan & McDermott, 2007). Finally, our results are consistent with previous research that has demonstrated a relationship between retrieval demands and testing effects (i.e., Halamish & Bjork, 2011). Though effects of repeated testing were present, the final retention differences across groups were smaller in this Experiment as compared to Experiment 1.

Experiment 3

The goals of Experiment 3 were to further explore the role of testing and restudy on associative binding of the stimuli. Previous research suggests that retrieval practice does promote binding. For example in a paired-associates retrieval practice experiment, Carpenter et al. (2006) found that retrieval practice resulted in better performance on a paired-associates cued recall test as compared to restudy even when the cue at test had previously been the target at study. Specifically, if participants had studied a series of A-B cue-target word pairs, at test, they could have been prompted with either 'A' or 'B' and asked to retrieve the associated word. Carpenter et al. found that even when tested with the 'B' half of the pair, retrieval practice improved memory performance compared to study alone. Research also has demonstrated that retrieval practice promoted integration of complex material. For example, participants demonstrated better learning of a related passage if they had first taken a test on a previously presented passage as compared to participants who had not been tested (Wissman, Rawson, & Pyc, 2011).

The use of triads allows us to vary the cue (the first, second, or third word in the triad) to determine how well the information has been bound. If the information has been well integrated, the use of a different cue should be less disruptive than when information is less tightly bound. In addition, a well-integrated memory should provide multiple retrieval paths (Tulving, 1974). These multiple paths (or cues) should facilitate better access to the target information. We predicted that retrieval would promote integration such that any part of the triad could serve as an effective cue. Binding should be strongest in situations where participants engaged in retrieval practice prior to restudy. However, if retrieval, in and of itself, is the critical factor in associative binding then we would expect to see equivalent performance between the STS and the STF groups. Retrieval practice should instantiate binding of at least part of the triad as well as highlight what information has not been successfully integrated. Restudy then should be targeted to the poorly bound information. In addition, restudy should further strengthen the binding established during an initial study episode.

Method

Participants

Sixty undergraduate students from the Richard Stockton College of New Jersey participated in this experiment for course research credit. Participants were equally distributed across the three groups.

Materials and procedure

The experimental design and procedure was identical to the previous experiments, with two exceptions. First, participants studied only nine related and nine unrelated triad groups. We reduced the number of total triads in order to raise performance on a test format that was expected to be challenging. Second, participants were given a varying-cue cued recall test. Participants studied groups of three related words. After initial study, two groups of participants received the varying-cue cued recall test and a third group engaged in repeated study.

The nature of the cued recall test was such that the cue varied for a given triad grouping. This test was designed to further examine associative binding and triad integration. The cue used on the cued recall tests shifted such that for each word in the triad served as a cue on the test. For example, for the triad DIAMOND RUBY EMERALD, participants could have received any of these three words as the cue at test. Specifically, if EMERALD was the cue, the subject would be presented with EMERALD in the centre

of the screen and would be required to produce the two associated words. For the Initial Test, a third of the triads were tested using the first word of the triad group as the cue. One-third used the second word as the cue. The last third used the third word as the cue.

For example, if participants were given the first word of the triad group as a cue on the Initial Test, they could have been given the first, second, or third word of the group on the Final Test. After the Initial Test or second study, participants either engaged in an additional study phase, or a filler task. Participants were told that a final test would occur 48 h later, in which they would be presented with a word from the triad group and would have to recall the two accompanying words. Cues varied on the Final Test as well. For one-third of the pairs, participants were given the same cue as was given on the first test. Thus, for these triads, participants were given the same cue that could have been in one of three positions for both the Initial and Final test. Cues between Initial and Final test systematically varied. The stimuli were counterbalanced such that all words in a triad served equally as a cue at test.

Results

Due to extremely low performance ($M = .03$), we eliminated unrelated triads from further analysis. As in Experiment 1, we scored the Initial Test cued recall using liberal and strict scoring criteria. As Table 3 demonstrates, Practice group had no effect on cued recall performance regardless of scoring criteria, Fs < 1. However, cued recall performance was affected by Cue. When scored using the liberal criterion, we found a marginal main effect of Cue, $F(1, 37) = 3.73$, $p = .06$, $\eta_p^2 = .09$. Under the strict criterion, the effect of Cue was significant, $F(1, 37) = 5.06$, $\eta_p^2 = .12$. Pairwise comparisons using a Bonferroni correction resulted in significant differences in performance between Cue 1 ($M = .15$) and Cue 3 ($M = .04$), $t(39) = 2.72$, $d = .55$, and between Cue 2 ($M = .15$) and Cue 3, $t(39) = 3.24$, $d = .68$. The difference between Cue 1 and Cue 2 was not significant, $t < 1$.

When examining performance on the Final Test using a 3 (Practice: SSS, STS, STF) × 2 (Test Cue Match: Same, Different) ANOVA, we found a main effect of Practice group, $F(1, 57) = 5.71$, $\eta_p^2 = .17$. Pairwise comparisons using a Bonferroni correction revealed a significant difference between the STS group ($M = .40$) and the STF group ($M = .18$), $t(38) = 3.33$, $d = 1.10$. The difference between the STF group and SSS group ($M = .24$) did not reach statistical significance. The interaction between Cue Match and Practice group was not significant, $F < 1$. A similar analysis using average performance scored under the strict criterion found a similar pattern of results. That is, we found a main effect of Practice group, $F(2, 57) = 6.05$, $\eta_p^2 = .18$. The difference between the SSS group ($M = .06$) and the STF group ($M = .07$) were not statistically significant, $t < 1$. However, the difference between the STF and STS groups ($M = .17$) were significant after a Bonferroni correction, $t(38) = 2.48$, $d = .79$. Similarly, the difference between the STS and SSS groups reached statistical significance, $t(38) = 2.57$, $d = .80$. No other effects were significant.

Table 3. Experiment 3—Cued recall performance on Final test as a function of cue congruency between Initial and Final test.

	Overall	Same cue	Different cue
Final test liberal			
SSS	.20 (.05)		
STS	.40 (.07)	.50 (.07)	.36 (.07)
STF	.18 (.05)	.28 (.07)	.13 (.03)
Final test strict			
SSS	.06 (.03)		
STS	.17 (.04)	.13 (.04)	.18 (.04)
STF	.07 (.04)	.10 (.05)	.05 (.02)

Note: Standard error in parentheses.

General discussion

The present results suggest that interpolated study and test produced better memory compared to either study or testing in isolation. This was true across three different test types, and across different scoring criteria. In Experiment 1, interpolated study and test led to the best performance on a cued recall test. In Experiment 2, interpolated study and test led to the lowest false alarm responding on an associative recognition test. Finally, in Experiment 3, interpolated study and test led to the best memory performance on a cued recall test when the test cue systematically shifted across all elements of a given triad. We believe that these effects are a largely a result of improved integration of the restudied material following a test.

The effects of repeated study

Study is clearly an important part of the learning process. However, study in isolation only has modest effects on the binding of information. When measuring memory via cued recall tests (Experiments 1 & 3), repeated study led to the poorest performance on the Final test. Notably, in Experiment 3, when the cue changed position, participants in the repeated study condition were near floor in their performance. When the test required flexible access to different parts of the triad, participants were clearly not able to access that information. In the context of recognition (Experiment 2), repeated study produced robust hit rates. When the triad was presented intact, participants were able to correctly recognise it at a high rate. The deficits in the repeated study condition became apparent in false alarms. Participants in the repeated study group demonstrated higher level of false alarm rates on average, than when study and testing were interpolated. We interpreted these results as demonstrating familiarity-based recognition responding and poor integration of the triad in the repeated study group.

The higher rates in false alarms in the SSS and STF groups as compared to the STS group suggest that

without interpolated study and test, participants are more prone to familiarity-based recognition decisions, as opposed to recollection-based decisions (e.g., Chan & McDermott, 2007). When recollective (more effortful) processes were required to discriminate between new and old elements of the triad, performance declined. Further, these recognition results are consistent with the results from the cued recall tests supporting the idea that tests that require recollective search requires a well-integrated memorial representation.

The effects of repeated testing

The present experiments also allowed for the examination of the effects of testing, in isolation, on the binding of the material. In Experiment 1, the data reflect the standard testing effect in both strict and liberal scoring. That is, participants who took an intervening test performed better on the Final test than those who repeatedly studied. These results suggest that repeated testing, even in isolation, promotes binding (see also Carpenter et al., 2006). Recognition performance in Experiment 2 is also suggestive of differential integration as a result of retrieval practice. That is, participants in the isolated testing group demonstrated on average lower false alarm rates as compared to participants who engaged in isolated study. That being said, isolated study led to higher hit rates than isolated testing. This was true for both related and unrelated lures. Only when study was interpolated with testing were false alarm rates low and hit rates high.

Two important conclusions emerge from these findings: (1) retrieval practice, even in isolation, promotes binding, and (2) study following retrieval practice may be targeted to less well learned information, further facilitating binding across multiple items. While the results of Experiments 1 and 2 suggest that repeated retrieval may promote binding, performance between the SSS and STF groups did not differ in Experiment 3. Recall that in Experiment 3 the test cue systematically shifted across all elements of a triad. In this experiment, the test cue shifted on the Initial test as well as the Final test. Similar previous research (Carpenter et al., 2006) utilised only 'A' as a cue from the A-B word pair on their Initial test. That is, on the first test participants were given what might be considered the default cue, or 'A'. Getting the default cue on the first test may have capitalised on principles of encoding specificity (cf. Tulving & Thomson, 1973) resulting in better Initial Test performance. Research suggests that one important requirement for the testing effect is to have successful performance on the Initial Test (Halamish & Bjork, 2011). When participants in our study were prompted with atypical cues, ('B' or 'C'), they may have been given less of an opportunity to use retrieval practice for initial binding. This may have resulted in the limited testing effect demonstrated by the STF group.

Given previous research demonstrating the superiority of testing over restudy (Karpicke & Roediger, 2008), one might ask why our isolated testing group did not perform better. Specifically, why did the STS group consistently outperform the STF group? We would argue that this has to do with the level of learning (and most likely binding) at the time that the tests were initiated. The aforementioned research had participants learn the material to a specific criterion before initiating retrieval practice. Even in situations where researchers have examined the benefits of testing prior to study (Grimaldi & Karpicke, 2012), these benefits only occur for word pairs that had relatively strong a priori associations. We believe that if we equated the overall level of learning prior to initial testing the differences between the STS group and the STF groups would be greatly minimised. Similarly, in Experiment 3, where the demands of the test were very challenging, performance for unrelated triads fell to floor. Coupled with previous research, the present results suggest that in order for testing effects to emerge, there must be some baseline level of pre-existing associations or participants may need to learn new associations to a specific criterion. In other words, the information has to be sufficiently encoded before testing can become maximally effective.

The effects of test potentiated study

While the present experiments allowed for the examination of studying and testing in isolation on binding, the primary focus of the manipulations was to examine how these processes worked in tandem. The results strongly suggest that study following a test aided the binding of material. In Experiment 1, with both liberal and strict scoring, the test potentiated study condition (STS) dramatically outperformed the other two conditions. Further the STS group demonstrated greater gains between the Initial and Final test than the STF group, suggesting the testing followed by study was more likely to promote new learning. In Experiment 2, the test potentiated study group demonstrated both the highest hit rate as well as the lowest false alarm rates. This suggests that not only did the restudy following a test lead to better memory for the individual units (as evident in the high hit rate in the repeated study condition) but it also allowed participants to successfully reject other previously studied items much more successfully. We interpreted this pattern as demonstrating improved associative binding of the triads in this condition. In Experiment 3, the test potentiated study group overall dramatically outperformed the other groups. But, more crucial to the focus of the present study is how relatively unaffected the participants in this group were when the cue shifted. These participants seemed to have the ability to more easily and flexibly access the entire triad regardless of which element they were given as a cue. The strong flexible access to information seems to override principles of encoding specificity (i.e., Tulving & Thomson, 1973). That is, after engaging in interpolated study and test, participants were better able

to recall correct elements of a triad regardless of whether the cue used at test matched that used during study.

Taken together, the results from these three experiments suggest that retrieval practice paired with restudy fosters successful binding as compared to conditions in which testing and restudy are not paired. Much recent research has been conducted to better understand the nature of the testing effect. Testing appears to lead to several downstream effortful cognitive processes. The first involves the initial test itself. Testing is an effortful process that leads to a broad search of memory. Previous research (Carpenter, 2011) has demonstrated that elaborative retrieval activates and makes accessible related information at the time of test. The second aspect of this process involves identifying and allocating attention to gaps in one's knowledge (Izawa, 1966). Related research (Gordon & Thomas, 2014; Gordon et al., 2015) has shown that taking a test identifies discrepancies in memory and aids in the processing of new related information. The final process in this sequence is test potentiated learning.

Finally, our results suggest that to reap the maximum benefit from test-potentiation effects, the material must already have some baseline level of integration. This is clearly demonstrated by the interaction found between practice group and triad type in Experiment 1. Although interpolated study and testing resulted in better final performance as compared to the other practice groups, the difference in final performance was significantly greater for related than unrelated pairs. Similarly, in Experiment 2, the hits for participants in the repeated test condition for unrelated pairs were the lowest recorded in this study, indicating the need for some baseline level of associations to produce a robust testing effect.

Conclusion

The present research primarily focused on these potentiation effects. We demonstrated that testing followed by restudy (1) improved overall memory performance by facilitating binding, (2) reduced false alarms driven by familiarity-based responding, and (3) allowed for better performance when different retrieval cues were utilised in the final test. We believe that the present research adds to this body of knowledge on the testing effect, and more importantly, helps elucidate previous underspecified mechanisms that may underlie test-potentiated learning.

Acknowledgements

The authors would like to thank John Dunlosky for his assistance in the initial planning of this project.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 940–945.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1118–1133.

Carpenter, S., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830.

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

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*, 153–170.

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

Chan, J. C. K., Wilford, M. M., & Hughes, K. L. (2012). Retrieval can increase or decrease suggestibility depending on how memory is tested: The importance of source complexity. *Journal of Memory and Language*, *67*, 78–85.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*, 369–377.

Gordon, L. T., Bulevich, J. B., & Thomas, A. K. (2015). Looking for answers in all the wrong places: How testing facilitates learning of misinformation. *Journal of Memory and Language*, *83*, 140–151.

Gordon, L. T., & Thomas, A. K. (2014). Testing potentiates new learning in the misinformation paradigm. *Memory & Cognition*, *42*, 186–197.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 801–812.

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, *56*(2), 212–242.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, *18*, 879–919.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, *83*, 340–344.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*, 200–224.

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

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, *38*, 116–124.

Kintsch, E., & Kintsch, W. (1978). The role of schemata in text comprehension. *International Journal of Psycholinguistic Research*, *52*, 17–29.

Kintsch, W., & Greene, E. (1978). The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes*, *1*, 1–13.

Lewis, M. R., & Mensink, M. C. (2012). Prereading questions and online text processing. *Discourse Processes*, *49*(5), 367–390.

McDaniel, M. A., Howard, D., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*, 516–522.

Naveh-Benjamin, M. (2000). Adult-age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*, 1170–1187.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved from http://www.usf.edu/FreeAssociation/

O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*, 389–397.

Roediger, H. L., & Butler, A. C. (2013). Retrieval practice (testing) effect. In H. L. Pashler (Ed.), *Encyclopedia of the mind* (pp. 660–661). Los Angeles, CA: Sage.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Roediger, H. L., & Nestojko, J. F. (2015). The relative benefits of studying and testing on long-term retention. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, & M. Styvers (Eds.), *Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin* (pp. 99–111). New York: Psychology Press.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1392–1399.

Thomas, A. K., Bulevich, J. B., & Chan, J. C. K. (2010). Reducing retrieval enhanced suggestibility through warning. *Journal of Memory & Language*, *63*, 149–157.

Thompson, C. P., Wegner, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall. A reappraisal. *Journal of Experiment Psychology: Human Learning and Memory*, *4*, 210–221.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*, 252–257.

Tulving, E. (1974). Cue-Dependent forgetting. *American Scientist*, *62*(1), 74–82.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior*, *13*, 181–193.

Van Overshelde, J., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An expanded and updated version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*, 289–334.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*, 1140–1147.

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does trying, and failing, to predict to-be-learned responses enhance later recall of those responses? *Memory & Cognition*, *42*, 1373–1383.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–1008.