



## Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention

Justin B. Knight<sup>a,\*</sup>, B. Hunter Ball<sup>a</sup>, Gene A. Brewer<sup>b</sup>, Michael R. DeWitt<sup>a</sup>, Richard L. Marsh<sup>a</sup>

<sup>a</sup> Department of Psychology, University of Georgia, Athens, GA, United States

<sup>b</sup> Department of Psychology, Arizona State University, Tempe, AZ, United States

### ARTICLE INFO

#### Article history:

Received 6 July 2011

revision received 14 December 2011

Available online 10 January 2012

#### Keywords:

Unsuccessful retrieval attempt

Error

Testing effect

Semantic activation

Mediator effectiveness

Retrieval practice

### ABSTRACT

Five experiments were conducted to examine how unsuccessful retrieval influences learning and subsequent memory. We used a cued-recall paradigm that produces many unsuccessful retrieval attempts (followed by feedback) and allows comparisons to be made between later memory for these trials and trials that only required reading or studying the pairs. On *read* trials participants studied cue–target pairs that were either weakly associated (DOOR–EXIT) or unrelated but identical in length (DOOR–SHOE). On *test* trials participants were given only the cue (either without [Exps. 1–3] or with [Exps. 4–5] prior experience with the pair items) and asked to guess the target which they were told was either semantically related or identical in length to the cue; then they received the correct cue–target pair to study. Unsuccessful retrieval attempts (i.e., guessing) relative to studying benefited retention for weakly associated pairs but impaired retention for unrelated pairs. This pattern of results occurred regardless of study duration (Experiments 1A and 1B), level of processing of the cue (Experiment 2), whether relatedness was manipulated between or within subjects (Experiment 5), and when guessing involved episodic as opposed to semantic retrieval (Experiments 4 and 5). However, this pattern was partly mediated by the ability to retrieve incorrect guesses during a final cued-recall test which may provide a link between the cue and target (Experiment 3). The current study demonstrates that unsuccessful retrieval attempts with immediate feedback not only enhance, but also can impair learning. This effect is robust and depends on elaborative semantic activation related to the answer and the effectiveness of incorrect guesses as mediating cues.

© 2011 Elsevier Inc. All rights reserved.

### Introduction

Testing memory has repeatedly been found to benefit learning of information relative to restudying the information (Roediger & Karpicke, 2006). The benefits of testing—termed the testing effect—suggest that the successful retrieval of material on some intervening test establishes a more durable memory of that material and leads to greater retention of that information. The testing effect has been found using a variety of materials including

standard word pairs (Carpenter, 2009), face–name pairs (Carpenter & DeLosh, 2005), foreign language vocabulary (Karpicke & Roediger, 2008), conceptual category exemplars (Jacoby, Wahlheim, & Coane, 2010), general knowledge facts (Carpenter, Pashler, & Cepeda, 2009), SAT II questions (Marsh, Agarwal, & Roediger, 2009), and text passages (Chan, 2010). The applicability across a wide range of domains has led a number of researchers to suggest that more tests should be incorporated into educational curriculums (Roediger & Karpicke, 2006; Rohrer & Pashler, 2010). Evidence demonstrating the benefits of testing in the classroom is beginning to accumulate (McDaniel, Agarwal, Huelser, McDerrott, & Roediger, 2011; McDaniel, Anderson, Derbish, & Morrisette, 2007;

\* Corresponding author. Address: Department of Psychology, University of Georgia, Athens, GA 30602, United States. Fax: +1 706 542 3275.

E-mail address: [jbknight@uga.edu](mailto:jbknight@uga.edu) (J.B. Knight).

McDaniel, Roediger, & McDermott, 2007; Rohrer & Pashler, 2010).

These findings demonstrate that tests clearly can serve as an additional learning opportunity. Nevertheless, when taking tests, errors inevitably occur; that is, the correct answer is unable to be retrieved or a wrong answer is incorrectly retrieved. The impact of these unsuccessful retrievals or tests (which will be used synonymously herein) on learning has received much less focus relative to the rich literature investigating successful retrieval attempts (for a review see Roediger & Butler, 2011). Accordingly, the benefits of testing have typically been presumed to occur only when retrieval was initially successful, though recent evidence suggests when feedback is provided that benefits to learning may accrue even on unsuccessful retrieval attempts (Kornell, Hays, & Bjork, 2009). The goals of the present study were to further test the proposal that unsuccessful tests can serve as learning experiences and to examine possible mechanisms and explanations that could account for the influence of unsuccessful retrieval attempts on memory acquisition.

Investigating the impact of errors on tests is an important endeavor, as many educators are wary of incorporating tests in place of other types of instruction because of their concern that the erroneous answers incorrectly produced by students may have negative impacts (Pashler et al., 2007). A better understanding of the influence of unsuccessful retrieval attempts could help alleviate some concerns about increasing the number of tests in the classroom and could provide guidance to ensure that tests are set up and administered in a way that is maximally beneficial. Furthermore, extant literature provides divergent predictions about the influence of unsuccessful tests. The errorless learning tradition holds that errors should be avoided during learning (Guthrie, 1952). This notion emanates from the proposal that when an error is produced it establishes an incorrect stimulus–response association, and thus, if an error occurs, it is likely to impair learning and lead to future errors. The support for this view has largely been derived from research on animal learning (e.g., Skinner, 1958; Terrace, 1963). However, errors on an initial test have been found to persist on future tests in human studies involving cued recall (Butler & Peterson, 1965), free recall (Roediger & Payne, 1982), and multiple choice tests (Fazio, Agarwal, Marsh, & Roediger, 2010; Marsh, Roediger, Bjork, & Bjork, 2007; Roediger & Marsh, 2005).

In an early study, Cunningham and Anderson (1968) had participants guess the target consonant trigram that would be paired with a two-digit number cue, and after they output their guess (which was typically incorrect) participants were presented with the correct number–trigram pair. The results revealed that on a subsequent cued-recall test participants in the guessing group were significantly less likely to recall the correct trigram relative to a group that initially read the pairs intact. They proposed that the production of the guess increased proactive interference (Postman & Underwood, 1973), and thus, hindered recall of the correct answer. This study provides direct evidence of the potential deleterious effects of errors and supports the claim that for optimal learning on tests to occur errors should be minimized. Furthermore, the error

minimization approach to learning has been successfully applied in rehabilitation for patients with lesions and neuropsychological disorders (Clare & Jones, 2008; Mueller, Palkovic, & Maynard, 2007).

Nevertheless, evidence also exists which demonstrates that errors during tests may not impede learning (Pashler, Rohrer, Cepeda, & Carpenter, 2007) and in some cases can facilitate learning (Kornell et al., 2009; Richland, Kornell, & Kao, 2009). Initial evidence for the potential beneficial role of unsuccessful tests was reported by Izawa (1967, 1970). Izawa demonstrated that multiple unsuccessful cued-recall tests without feedback (e.g. 5 tests), relative to fewer tests, led to greater learning on a subsequent intact presentation. Kane and Anderson (1983) had participants guess the final word of undetermined sentences that could each be completed with a variety of words and found that memory for those sentences was superior to that of intact sentences which had simply been read, even though the majority of their initial guesses were incorrect. With this work, however, there remained the possibility that the small proportion of sentences for which participants guessed correctly could have led to the better memory performance relative to the read sentences. A recent study removed the influence of correct guesses by (i) using weakly associated pairs, (ii) having participants guess the target before being presented with the intact cue and target, and (iii) only analyzing final cued-recall performance for the items that were initially incorrectly guessed (Kornell et al., 2009). Cued-recall performance on a later test was greater for the pairs that were unsuccessfully tested as compared to pairs that were simply read for the same amount of time. This effect was present on tests that occurred after five minutes and more than 24 h. Thus, these results demonstrated that, in some cases, even unsuccessful retrieval attempts can serve as effective learning opportunities; however, the mechanism(s) operating to support the beneficial role of unsuccessful tests remain(s) unidentified. Further, the conditions that determine whether unsuccessful tests will facilitate or impair learning have yet to be fully characterized.

Several mechanisms have been proposed that may underlie the enhancement of learning through unsuccessful retrieval. One explanation was derived from the long research tradition demonstrating that the type or level of processing during encoding plays a determining role in learning ( Craik & Lockhart, 1972); it also follows from proposals that successful testing benefits memory through elaborative activation during retrieval practice (Carpenter, 2009; Pyc & Rawson, 2009). Accordingly, unsuccessful retrieval may encourage deep, semantic processing during encoding of the pair (Kornell et al., 2009; Richland et al., 2009). That is, when presented with a cue, the participant queries memory to guess the target and likely samples potential responses which activates a related semantic network (e.g., Anderson, 1983). This process encourages deep processing of the cue and activates a network conceptually related to the target which would likely establish a fertile context for encoding the intact pair. For example, if presented with the word *nature*, participants may think of a tree, park, mountain, river, etc.; then when presented with the cue and correct target *nature-trail*—participants could imagine a park with

a nature trail running along a river with trees all around and a mountain in the distance. Thus, the unsuccessful test would facilitate elaborative processing with multiple links between the cue and target, and consequently, facilitate learning (e.g., Carpenter, 2009).

Additionally, unsuccessful retrieval could benefit learning if the generated guess served as a mediating cue between the experimenter-given cue and target on a future retrieval attempt (e.g., Pyc & Rawson, 2010; Soraci et al., 1999). That is, the information activated during the unsuccessful retrieval attempt (and hence during the encoding of the cue–target pair) could function as a retrieval cue to aid in the memory search on the final test. This explanation is highly similar to the recently proposed mediator effectiveness hypothesis (Pyc & Rawson, 2010). On an initial study trial of foreign language word pairs, Pyc and Rawson had participants produce a keyword mediator that was related to the pair and that was meant to help them remember the pair. Then participants either restudied the pairs or were tested on the pairs, and for each pair all participants continued to produce a mediator (it could be the same one as before or a new one). On a final test, they found that test trials, relative to study trials, led to both greater accessibility of the mediator and greater diagnosticity of that mediator in supporting retrieval of the target. Thus, they demonstrated that at least one way successful testing benefits memory is by facilitating the effectiveness of mediating cues between the cue (question) and target (answer). In the current study we will show that the extent to which the participant can dissociate between their generated guess and the experimenter-given target in the search set will influence whether memory is facilitated through a mediating cue or impaired through the build-up of proactive interference (Cunningham & Anderson, 1968; Postman & Underwood, 1973).

A third mechanism that could support a beneficial influence of unsuccessful retrieval attempts is that the incorrect guesses could lead to a suppression of the erroneous retrieval routes (e.g., Carrier & Pashler, 1992). This explanation could be viewed as the complement to the proposal that the testing effect (i.e., benefit of successful retrieval) results from the strengthening of the appropriate retrieval routes or the creation of multiple retrieval routes during the initial retrieval practice (Bjork, 1975; McDaniel & Masson, 1985). Thus, when an error is produced and corrected, the incorrect retrieval route may be suppressed which, in turn, increases the probability that the correct retrieval route will be accessed on the final test. By the incorrect retrieval route we are referring to the pathway of information linking the cue and the participant's incorrect guess, whereas the correct retrieval route refers to the link between the cue and the correct target. Further, this proposal is consistent with connectionist models of error-correction learning which have been applied to the testing effect previously (Carrier & Pashler, 1992; McClelland & Rumelhart, 1986). In this view, the association between the cue and target is learned by adjusting the connection weights between these elements based on the difference between the network (i.e., participant) generated response and the actual target.

Lastly, a fourth explanation would be that the unsuccessful retrieval attempt makes participants aware of the

difficulty of retrieving the desired target when presented with that particular cue which leads the participant to devote more attention and/or effort to encoding the subsequently presented intact cue–target pair (e.g., Chan & Langley, 2010). That is, the inability to retrieve the target may motivate participants to more attentively encode the pairs following unsuccessful tests which could support the boost to learning for those pairs (Craig, Govoni, Naveh-Benjamin, & Anderson, 1996).

The potential mechanisms described above are likely not exhaustive nor are they proposed to be mutually exclusive; it is certainly possible that multiple mechanisms may support the influence that unsuccessful retrieval attempts have on learning. A complete disambiguation of these explanatory mechanisms would be too lofty a goal for the current study. Rather we aim to provide an initial examination of the relative contributions of these processes to unsuccessful tests in order to better understand their benefits and determine if and when unsuccessful tests may be harmful to learning.

## Experiment 1A

The basic trial design of this experiment was adapted from Experiment 4 of Kornell et al. (2009). There were three phases: study, delay, and test. During study, all participants completed 60 trials, with half of those being read trials and half being test trials. On read trials a cue–target pair was presented intact on the screen for 13 s. On test trials the cue word was presented alone for 8 s, and the participants were told to try to guess the target word that would be paired with it and enter their response using the keyboard. Subsequently, the cue and target were presented together for 5 s. Therefore, the total time experiencing the cues was equated for the read and test trials. This experiment was concerned with examining the extent to which the activation of a related semantic network would contribute to the influence of unsuccessful retrieval attempts on learning. Thus, the semantic association between the cue and target of a pair was varied between two groups of participants. Each cue and target in a pair in one group were weakly associated, similar to the stimuli used by Kornell et al. (2009), whereas the cue and target in a given pair in the other group were unrelated and matched on the number of letters. Therefore, if the activation of a related semantic network during an unsuccessful retrieval attempt supports the facilitation to learning, then no benefits on the final test should be observed for test trials relative to read trials in the group with pairs of unrelated cues and targets. However, if an increase in attention to encoding the pair following an unsuccessful test supports the benefit to learning, then a facilitation to tested pairs relative to read pairs should result for both groups on the final test.

### Method

#### Participants

Undergraduates from the University of Georgia participated in exchange for partial credit toward a class research

requirement. Each participant was tested individually in sessions that lasted approximately 30 min. Participants ( $N = 60$ ) were randomly assigned to one of two conditions. We used a  $2 \times 2$  mixed design with factors of study type (read or test) which was manipulated within subjects and relatedness (unrelated or related) which was manipulated between subjects.

### Materials

In the unrelated group, the material consisted of 60 unassociated word pairs, in which the cue and target of each pair had the same number of letters (e.g. *DOOR–SALT*). The cue and target were matched on the orthographic feature of letter count. Half of the word pairs were randomly assigned to either the read or the test condition (mean KF frequency of 44.513). Half of the pairs in each condition consisted of a cue and target that were each four letters long, and the other half were five letters. The words were pseudo-randomly paired. Pairs were formed so that the cue and target were not associated to each other or to any other word on the list. This was done by ensuring that for each item, none of the other items on the list were output as an associate in the (Nelson, McEvoy, & Schreiber, 1998) norms.

The study list for the related group was created in a similar manner, with the exception that the word pairs consisted of 60 weakly associated word pairs. As in the unrelated group, half the pairs were assigned to the read trials and half to the test trials. For each pair the cue had a weak forward association to the target word that was no greater than .05 ( $M = .03$ ,  $SD = .016$ ) based on the (Nelson et al., 1998) norms (e.g. *DOOR–EXIT*, *NIGHT–STAR*; Carpenter, Pashler, & Vul, 2006; Kornell et al., 2009). Thus, the pairs were semantically related; however, the targets would rarely be guessed correctly (i.e., only 5% of people in Nelson et al.'s studies output these targets as their first associate). The cue words had no forward association to any other words on the list except its paired target. Half of the cues in each condition were four letters long, and the other half were five letters long. The length of the cues across groups was identical. The number of letters in the target words was matched as closely as possible to the number in the cues. The target word length ranged from 4 to 7 letters ( $M = 4.7$ ) and the mean word frequency of the cues and targets was 43.026.

### Procedure

In the study phase, half of the word pairs were presented in the read condition, and half were presented in the test condition. In the read condition, a cue and target were presented together for 13 s, and participants were informed to study the pair of words for a later cued-recall memory test. On the test trials, the cue appeared with question marks beside it for eight seconds, and the cue and target were presented for five seconds. Thus, the total time that the cues were experienced was equated between the read and the test condition (i.e., 13 s). For test trials, participants tried to guess the target and type in their response (which appeared on the screen below the cue) before eight seconds elapsed. In the unrelated group, participants were told that the target would have the same

number of letters as cue, thus they should guess a word with the same number of letters. In the related condition, participants were told that target would be semantically related to the cue, thus they should guess a word that is associated with the cue. In both conditions, they were told they had 8 s to type in a target before the cue would appear with its correct target. At that point they were to study the pair for a later cued-recall memory test. The 60 word pairs were presented randomly for each participant. After the study list, participants completed a distractor task in which they had five minutes to type as many names of countries as they could.

The final phase consisted of a cued-recall test in which participants were to recall the correct target for each cue from the study phase. Each of the 60 cues was presented randomly one at a time in the center of the computer screen. Participants typed their response and then pressed the enter key to see the next cue. The participants had the option to type "XXX" and pass onto the next trial if they absolutely could not remember the target.

### Results and discussion

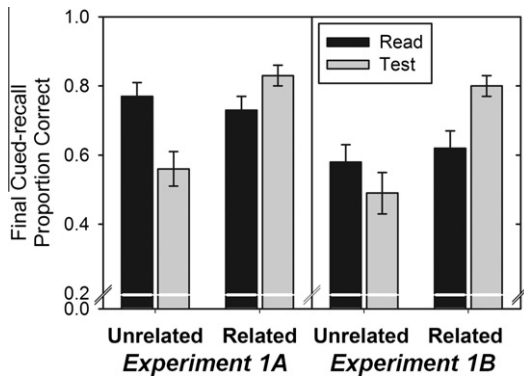
During the study phase participants guessed the correct target on 0 and <1% of the test trials in the unrelated and related conditions, respectively; and this was the case for all subsequent experiments (except Exp. 2 in which targets were guessed 0% and 2%, respectively, and except Exp. 5 which will be discussed below). Thus, correct retrievals happened rarely, and items that were correctly guessed here and in all other experiments were not included in further analysis as our interest was in the effects of unsuccessful retrieval attempts on learning. During the study phase, every participant generated a guess on some proportion of the trials and this occurred on >89% of all test trials for both conditions in all experiments. Thus, unsuccessful tests with errors of commission, as opposed to omission, occurred on the majority of test trials in all experiments.<sup>1</sup> The high percentages of commission errors suggest participants were trying to retrieve or guess the target on a vast majority of the test trials for both conditions across all experiments. For analyses in this and all following experiments, final test means for each response type were computed as the proportion of the number of responses of each type relative to the total number of responses output.<sup>2</sup>

Refer to Fig. 1 for plotted means of final cued-recall performance for Experiments 1A and 1B. A  $2 \times 2$  mixed model ANOVA that had a within-subjects factor of study (read vs test) and a between-subjects factor of relatedness (unrelated vs related) was computed. Significant main effects of study and relatedness were found  $F(1, 58) = 5.16$ ,  $p = .027$ ,  $\eta_p^2 = .082$ ,  $F(1, 58) = 5.04$ ,  $p = .029$ ,  $\eta_p^2 = .08$ , respectively, with the read items and the related condition

<sup>1</sup> The reported pattern of results was observed for analyses restricted to items that received either an error of commission or omission on the initial study/test phase.

<sup>2</sup> Final test trials on which participants passed and did not enter a response were not included in the reported analyses. Analyses conducted on means computed as the proportion of a given response type to the total number of items produced similar patterns of results as those reported.





**Fig. 1.** Final cued-recall performance for Experiments 1A and 1B. Error bars reflect SEM.

exhibiting better cued-recall accuracy. These effects were qualified by a significant interaction,  $F(1, 58) = 38.55$ ,  $p < .001$ ,  $\eta_p^2 = .399$ . Follow-up comparisons indicated that the interaction was driven by lower cued-recall accuracy for the test pairs relative to the read pairs in the unrelated condition  $t(29) = 5.14$ ,  $p < .001$ ,  $d = .86$ , whereas better cued-recall accuracy was observed for the test pairs relative to the read pairs in the related condition,  $t(29) = -3.49$ ,  $p = .002$ ,  $d = .51$  (see Fig. 1).

Thus, being tested on a pair and unsuccessfully retrieving the target helped later memory if the cue and target were associated, yet it impaired later memory if the cue and target were unrelated and only shared the same number of letters. The benefit of unsuccessful retrieval in the related condition replicated Kornell et al. (2009), and it demonstrates the powerful influence of unsuccessful retrieval attempts, as the intact cue–target pairs were actually presented for longer on the read trials (13 s) than the test trials (5 s). Only observing a benefit for the related pairs, and not the unrelated pairs, demonstrates that the activation of a semantic network may be involved in the facilitative effects of unsuccessful retrieval attempts. Incorrectly retrieving the target was beneficial to the extent that a semantic network that overlaps with the cue and target could be activated during the unsuccessful test.

To ascertain the strategies participants used on unsuccessful retrieval attempts, we had participants complete

**Table 1**

Percentages of participants' responses for characteristics relied upon for unsuccessful retrieval attempts in Experiments 1 and 2.

	Characteristics	
	Orthographic/ phonologic (%)	Semantic (%)
<i>Experiment 1A</i>		
Unrelated	27	55
Related	5	82
<i>Experiment 1B</i>		
Unrelated	32	52
Related	4	83
<i>Experiment 2</i>		
Unrelated	3	97
Related	33	67

Note. Remaining percentage of participants reported relying on a combination of all three characteristics.

a post-experimental questionnaire on which they reported the characteristics or processes that they relied upon to generate guesses during the read-test phase in Experiments 1A, 1B, and 2. A blind rater rated participants' responses according to whether their strategy relied upon semantic, orthographic/phonologic, or some combination of those. The percentages of responses across participants falling within each category for Experiments 1A, 1B, and 2 are listed in Table 1. In Experiment 1A, more people in the unrelated than the related condition reported using orthographic or phonologic features to generate guesses for test pairs, whereas those in the related condition relied more upon semantic features than did those in the unrelated condition (Table 1). The remaining percentage of participants reported relying on a combination of all three characteristics. These reported strategies suggest that participants were trying to guess the tested targets in both conditions and demonstrate differential reliance upon orthographic/phonologic and semantic features for the unrelated relative to the related conditions. Interestingly, when the cue and the target were unassociated an unsuccessful retrieval attempt was actually harmful to later memory. Experiment 1B was aimed at determining if the shorter presentation time for intact pairs in the test relative to the read trials could account for this detriment.

## Experiment 1B

### Method

#### Participants

Undergraduates from the University of Georgia participated in exchange for partial credit toward a class research requirement. Each participant was tested individually in sessions that lasted approximately 30 min. Participants ( $N = 60$ ) were randomly assigned to one of two conditions. We used a  $2 \times 2$  mixed design with a within-subjects factor of study (read or test) and a between-subjects factor of relatedness (unrelated or related).

#### Materials

The materials used here were identical to those used in Experiment 1A.

#### Procedure

This experiment followed the same procedures as those in Experiment 1A, with one exception. Whereas in Experiment 1A the intact pair was presented for 13 s on the read trials and only 5 s on the test trials, here intact read and test pairs were both presented for 5 s. Thus, on test trials the cue was presented with a question mark beside it and participants were given eight seconds to try to guess the target, then they studied the correct pair for 5 s. On the read trials the cue and target were presented together for five seconds; this adjustment served to equate presentation time for the intact pairs across the read and test trials. Thus, any impairment of unsuccessful retrieval attempts observed in the unrelated condition could not be totally accounted for by differences in time spent with the intact pair. As in Experiment 1A, the relatedness of

the cue and target was manipulated between conditions. If the impairments observed for the unrelated test pairs relative to the read pairs were solely due to the longer presentation time for the read pairs, then this experiment should reveal no detriments for test pairs in the unrelated condition. However, if unsuccessful tests on unrelated pairs do systematically interfere with learning, then detriments should remain. Given test pairs in the related condition were remembered better than read trials in Experiment 1A, that benefit should remain or increase here.

### Results and discussion

A 2 study (read vs test)  $\times$  2 relatedness (unrelated vs related) mixed ANOVA examining cued-recall accuracy resulted in both a significant main effect of relatedness and a significant interaction between relatedness and study,  $F(1, 58) = 8.55, p = .005, \eta_p^2 = .128, F(1, 58) = 27.51, p < .001, \eta_p^2 = .32$ , respectively. The main effect of study was not significant,  $F(1, 58) = 3.64, p = .061$ . Follow-up comparisons investigating the interaction revealed that relative to the read pairs the unrelated condition exhibited lower cued-recall accuracy for the test pairs,  $t(29) = 2.12, p = .042, d = .3$ , whereas the related condition exhibited better cued-recall accuracy for the test pairs,  $t(29) = -5.77, p < .001, d = .9$ , (Fig. 1). Thus, even when the total time spent with the items was lower for the read pairs, unsuccessful retrieval attempts in the unrelated condition still impaired learning of the test pairs. Therefore, the detriment of unsuccessful tests when the cue and target are unrelated is not due to less time spent studying the intact pair; rather, unsuccessful retrieval attempts appear to be inherently deleterious in some cases. Regarding reported strategies, the same pattern from Experiment 1A was observed here. Participants reported a greater reliance upon orthographic/phonologic characteristics to generate guesses in the unrelated relative to the related condition and a greater reliance upon semantic characteristics in the related relative to the unrelated condition (Table 1). Overall, the results from this experiment replicated the findings of Experiment 1A illustrating that differences in total exposure time were not responsible for differences in cued-recall performance. Both the benefits and detriments to cued-recall performance were driven by unsuccessfully retrieving target information at acquisition.

### Experiment 2

The benefits in final cued-recall accuracy observed for the unsuccessfully tested related pairs are suggestive that the elaborative activation of a semantic network that overlaps with the cue and the subsequently presented target is a mechanism that supports the benefit of unsuccessful retrieval attempts. However, the relatedness of the cue–target pairs and the level of processing of the cue were not fully distinguishable in Experiments 1A and 1B. That is, participants in the related condition were required to process the cue's semantic meaning in order to guess a word that was associated with it, whereas those in the unrelated condition could simply process the number of letters that

the cue contained in order to guess another word that had the same number of letters. The potential of deeper processing of the cue in the related condition relative to the unrelated condition may have contributed to the benefits observed for unsuccessful tests in the related condition (Craig & Lockhart, 1972; Craig & Tulving, 1975). Experiment 2 was aimed at investigating the extent to which the facilitation of unsuccessful tests in the related group can be explained by a levels-of-processing effect on the cues.

### Method

#### Participants

Undergraduate students from the University of Georgia participated in exchange for partial credit toward a research requirement. Each person was tested individually in sessions that lasted approximately 30 min. Participants ( $N = 70$ ) were randomly assigned to one of the two conditions—unrelated and related.

#### Materials

The materials were identical to that of Experiment 1, except in the related condition, we equated the word length for the cues and targets, where each cue had the same number of letters as the target (KF 46.282, HAL log freq 9.282). The cues had a weak forward association (less than .05;  $M = .03, SD = .014$ ) to the targets.

#### Procedure

The procedure for Experiment 2 was identical to Experiment 1A with one exception. The instructions for the test trials were switched. In the unrelated condition, participants were told to guess a word that was *associated* to the cue when it appeared on the screen with question marks beside it. In the related condition, participants were told to guess a word that had the same number of *letters* as the cue. In both conditions, they were told to study the *correct* cue and target when it appeared on the screen after they typed in their guess. Importantly, the cue and the target in the unrelated condition were not related and only had the same number of the letters. The cue and the target in the related condition were semantically related. The presentation timing of read and test trials were identical to those used in Experiment 1A. With the change in instructions the cues in the unrelated condition would receive the deeper level of processing, whereas the cues in the related condition would be subjected to a shallower level of processing (Craig & Lockhart, 1972). Thus, if the effects of unsuccessful retrieval were due to levels-of-processing in Experiment 1, then cued-recall of the tested targets in the unrelated condition should be facilitated relative to the read trials. However, if cue–target semantic relatedness was the key contributor to the Experiment 1 findings, then the results of this experiment should be similar to those found in the first experiment.

### Results and discussion

Data for final cued-recall accuracy are plotted in Fig. 2 for this Experiment. A mixed ANOVA with a within

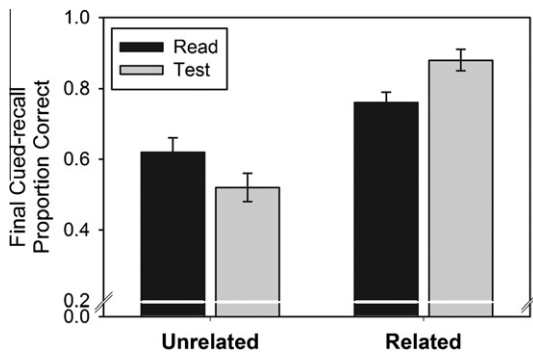


Fig. 2. Final cued-recall performance for Experiment 2. Error bars reflect SEM.

subjects factor of study (Read vs. Test) and a between subjects factor of relatedness (Unrelated vs. Related) was computed. The analysis revealed a significant main effect of condition,  $F(1, 68) = 30.73$ ,  $p < .001$ ,  $\eta_p^2 = .31$ , with the related condition exhibiting higher cued recall accuracy. This effect was qualified by a significant interaction,  $F(1, 68) = 24.63$ ,  $p < .001$ ,  $\eta_p^2 = .27$ . Follow-up comparisons revealed that the interaction resulted from the read items being better remembered in the unrelated condition,  $t(34) = 2.66$ ,  $p = .012$ ,  $d = .43$ , and the test items being better remembered in the related condition,  $t(34) = -5.12$ ,  $p < .001$ ,  $d = 1.17$ , (Fig. 2). This pattern of results replicates Experiment 1 and demonstrates that the benefit of unsuccessful retrieval attempts found in Experiment 1 for the related condition was likely not due to a deeper level of processing carried out on the cues in that condition. That is, in the present experiment the items in the unrelated condition were processed deeply at study in that participants were guessing a word that was semantically related to the cue, yet these unsuccessful retrievals still led to worse memory performance for the test relative to the read items. If the level of processing was the sole contributing factor to the influence of unsuccessful tests, memory accuracy for the unrelated test pairs (which received deep processing) should either be better than or, at the least, equal to accuracy for the unrelated read pairs, which was not observed here.

The results from Experiment 2 provide evidence that the benefit of unsuccessful retrieval to learning is not primarily driven by the level of processing of the cue, and this notion is additionally supported by the strategies participants reported using. In the unrelated relative to the related condition a greater number of participants reported reliance on semantic characteristics to generate guesses, whereas reported reliance on orthographic or phonologic characteristics displayed the opposite pattern between conditions (Table 1). Comparing the pattern of reported strategies used here to those in Experiments 1A and 1B reveals that the unrelated relative to the related condition reported engaging semantic processing more and orthographic or phonologic processing less in Experiment 2, whereas the opposite was observed in the initial experiments (see Table 1). These findings suggest that the levels-of-processing strategies engaged for the unrelated versus

the related condition differed between Experiments 1 and 2 yet the same pattern of final test performance was observed, further supporting the notion that these effects are not fully explained by a levels-of-processing account.<sup>3</sup> Additionally, the Experiment 2 results suggest there may be another mechanism by which unsuccessful retrieval facilitates memory in addition to the activation of an elaborative semantic network. Two possibilities are that the incorrectly generated guess may serve as a mediating cue between the cue and the target (Pyc & Rawson, 2010; Soraci et al., 1999), or the incorrect retrieval route (i.e., linking the cue and guess) may be suppressed and thus increases the likelihood that the correct retrieval route (i.e., between the cue and target) is accessed at the final test (Carrier & Pashler, 1992). These two possibilities make competing predictions about the accessibility of the incorrectly generated words from study; these predictions were tested in Experiment 3.

### Experiment 3

This experiment was aimed at elucidating if cue mediation or erroneous retrieval route suppression plays a role in the benefit of unsuccessful retrieval attempts to learning. The mediating cue hypothesis would predict that participants' generated guesses would be accessible at the final test and that the retrieved guess would serve as a mediator to help elicit the target. Pyc and Rawson (2010) termed the accessibility of the mediator, mediator retrieval, and its ability to elicit the target, mediator decoding. Contrary to this view, the proposal that the benefits of unsuccessful retrieval result from a suppression of incorrect retrieval routes (Carrier & Pashler, 1992; Kornell et al., 2009) would predict that participants' generated guesses should be largely inaccessible on the final test. This suppression proposal also predicts an inverse relationship between target and guess retrieval, in that successful retrieval of the target should be more likely when the incorrect retrieval route is suppressed and the incorrect guess is largely inaccessible. To arbitrate between these competing predictions, we asked participants on the final test to recall their generated guess for each pair, as well as the correct target. To the extent that guess accessibility (or lack thereof) is influential in the benefit to learning of unsuccessful retrievals, then this measure should differ for conditions in which unsuccessful retrieval is beneficial (e.g., related pairs) versus harmful (i.e., unrelated pairs). Since the primary interest of this experiment was participants' memory for their guess, we eliminated read trials and manipulated the relatedness of the pairs within subjects.

<sup>3</sup> The finding that some participants reported relying on semantic features when instructed to guess the target based on the length of the cue (i.e., orthographic feature) is not surprising considering people's natural tendency to process the meaning of information (i.e., overflow coding; Craik & Tulving, 1975). Nevertheless, a levels-of-processing explanation does not parsimoniously account for these data. That is, nearly all reports in the unrelated condition (96%) indicated participants relied on semantic characteristics to guess targets; yet final cued-recall performance for test pairs was still significantly impaired relative to read pairs. This finding is consistent with our proposal that the benefits of unsuccessful tests are in part dependent on whether the unsuccessful retrieval attempt activates a conceptual network related to the target (i.e., correct answer).

## Method

### Participants

Undergraduates from the University of Georgia participated in exchange for partial credit toward a research requirement. Each person was tested individually in sessions that lasted approximately 45 min. We used a within-subjects design ( $n = 30$ ) with two independent variables: relatedness (unrelated and related) and recall type (guessed item and target).

### Materials

The stimuli used were the same from Experiment 2. However, only the 30 unrelated test pairs and the 30 related test pairs were presented here. The study list was comprised of 60 word pairs, 30 were unrelated word pairs having only the same number of letters, and 30 were weakly associated word pairs ( $<.05$ ;  $M = .03$ ,  $SD = .012$ ).

### Procedure

All 60 trials were test trials. Participants received instructions to guess the target on each trial when a cue appeared. A prompt at the top of the screen would indicate whether the target had the same number of letters as the cue or whether it was semantically related to the cue. On the unrelated trials, a cue appeared on the screen with the prompt *LETTERS* above it. This indicated to the participants that they were to guess a word with the same number of letters as the cue. On the related trials, a cue appeared on the screen with the prompt *ASSOCIATE* above it, which indicated that participants should guess a word that would be associated with the cue. For both trial types, participants had 8 s to guess a word, and then the correct word pair appeared and was to be studied for a later cued-recall memory test. All 60 word pairs were randomly presented. Participants completed a practice phase before beginning the study phase. After the study phase was completed, participants engaged in the 5 min country generation distractor task as in the previous experiments. During the test phase, participants were given instructions for the cued-recall test. For each cue, participants were asked to recall the word they guessed earlier, when that cue was presented, and recall the correct target. When the cue appeared, they were to first type in the word they guessed earlier then press enter to record their response. Immediately following, they were presented with the same cue again and asked to type in the target that was paired with the cue previously. All 60 cues were randomly presented.

### Results and discussion

Refer to Fig. 3 for means of final cued-recall accuracy from this Experiment. Final test performance was initially analyzed using a repeated measures ANOVA with within-subjects factors of relatedness (unrelated vs. related) and recall type (generated vs. target). Results indicated that the main effect of relatedness was significant,  $F(1, 29) = 107.1$ ,  $p < .001$ ,  $\eta_p^2 = .79$ , indicating that overall cued-recall accuracy was greater for the related pairs. The main effect of recall type was also significant,  $F(1, 29) = 6.3$ ,  $p = .018$ ,  $\eta_p^2 = .18$ , as generated items

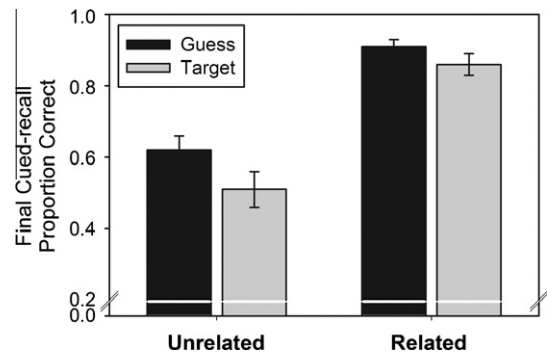


Fig. 3. Final cued-recall performance for participants' initial guesses and targets in Experiment 3. Error bars reflect SEM.

were better remembered than the target items. The interaction was not significant,  $F(1, 29) = 2.1$ ,  $p = .16$ . Further, for both types of pairs participants were able to recall their generated guess on a majority of the trials which suggests that the incorrect retrieval routes were not suppressed (Fig. 3). Although there were no read trial comparisons in this experiment, cued-recall accuracy for the tested unrelated and related pairs were similar to the performance observed in Experiments 1A and 1B, thus, unsuccessful tests here appear to have influenced memory similar to the other experiments. Follow-up comparisons revealed that participants successfully recalled significantly more generated guesses and more correct targets for the related pairs relative to the unrelated pairs,  $t(29) = -8.07$ ,  $p < .001$ ,  $d = 1.66$ ,  $t(29) = -8.75$ ,  $p < .001$ ,  $d = 1.52$ , respectively. Thus, the accessibility of the mediator and the target was greater for the related pairs compared to the unrelated pairs. Greater memory for items from related pairs than unrelated pairs may reflect an effect similar to the list strength effect (Ratcliff, Clark, & Shiffrin, 1990) whereby the increased strength of the related pairs led to competition and reduced the accessibility of the unrelated pairs (Verde, 2009).

For the words that were generated as a guess for unrelated and related pairs, we also sought to examine mediator decoding, that is, the extent to which greater generated guess retrieval is associated with greater target retrieval (Pyc & Rawson, 2010). For each participant, we computed the proportion of successfully retrieved targets on the final test that were preceded by retrieval of the generated guess for both unrelated and related pairs. Comparisons between unrelated and related pairs revealed that a significantly greater proportion of retrieved targets were preceded by retrieval of the generated guess for the related pairs (Related:  $M = .83$ ,  $SE = .04$  versus Unrelated:  $M = .44$ ,  $SE = .06$ ),  $t(29) = -6.45$ ,  $p < .001$ ,  $d = 1.48$ . Thus, mediator decoding was greater for the related pairs, suggesting a positive association between target retrieval and generated guess retrieval. This association was more robust for the related than the unrelated pairs.

One may contend that the generated guesses were simply more accessible for the related pairs, and that this inflated the assessment of the proportion of retrieved targets that was preceded by retrieved guesses. Therefore, we examined only the trials that participants retrieved



their previously generated guess. We computed the proportion of retrieved generated guesses that were followed by successful retrieval of the target. When only considering the guesses that participants were able to retrieve, direct comparisons between unrelated and related pairs revealed that this assessment of mediator decoding was significantly greater for the related pairs (Related:  $M = .64$ ,  $SE = .04$  versus Unrelated:  $M = .17$ ,  $SE = .03$ ),  $t(29) = -12.15$ ,  $p < .001$ ,  $d = 2.44$ . Thus, the majority of retrieved guesses were followed by successfully retrieved targets of the related but not the unrelated pairs, suggesting guesses may have facilitated the memory search for the target.

We chose an arbitrary cutoff of greater than 50% of guesses being retrieved to ascertain whether the link between the cue and the incorrect guesses was suppressed; therefore, we acknowledge that the incorrect retrieval route suppression account cannot be fully ruled out. However, the suppression account of the benefits of unsuccessful retrievals predicts that there should be a negative relationship between guess retrieval and target retrieval whereby worse guess retrieval should accompany better target retrieval. In contrast, here we found a clear positive relationship between guess and target retrieval in conditions that we have shown to produce beneficial influences of unsuccessful retrieval (i.e., with related pairs). Accordingly, the mediating cue hypothesis seems to provide a more parsimonious account of these data. The enhanced mediator retrieval and decoding for the related relative to the unrelated condition suggest that increasing the effectiveness of mediating information may be an influential mechanism in the benefits of unsuccessful retrieval attempts.

#### Experiment 4

All the experiments examining the influence of unsuccessful tests both here and in previous research (Kornell et al., 2009) have had participants try to generate or guess the target member on test trials without them having had previous episodic experience with the given cue–target pairs. Thus, evidence on the effects of unsuccessful retrieval attempts is currently restricted to the generation of information from one's semantic knowledge (i.e., words associated to a presented cue). Recent work investigating successful retrieval has demonstrated distinctions between the influences of episodic retrieval attempts versus semantic generation on retention (Karpicke & Zaromb, 2010). Further, the typical tests administered in educational environments are concerned with memory for information that one has experienced as a part of the class. Accordingly, it is important to determine if the observed influences of unsuccessful retrieval attempts extend to conditions that are more similar to educational tests—retrieval from a previous episodic experience. This issue was the focus of this experiment.

#### Method

##### Participants

Undergraduate students from the University of Georgia participated in exchange for partial credit toward a research requirement. Each person was tested individually

in sessions that lasted approximately 30 min. Participants ( $N = 60$ ) were randomly assigned to one of the two conditions—unrelated and related.

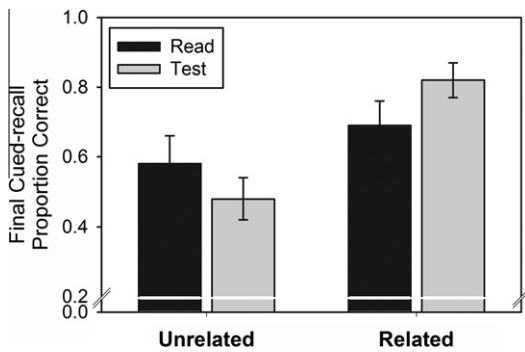
##### Materials

The same materials from Experiment 2 were used in this experiment.

##### Procedure

In the initial phase, 120 words were presented on the screen one at a time. Each word was presented for 1 s, and participants were instructed to read each word silently. Hence, during the first phase, participants were not aware of a future memory test. After completion of the first phase, participants completed the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970) and the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003) for an unrelated interest. Completion of these scales took about 5 min then the second phase began. The words that were presented in the first phase served as the cue–target pairs in the second phase. As in the previous experiments, the cue and target in each pair had the same number of letters and was unassociated in the unrelated condition, whereas the cue and target in each pair had the same number of letters and were associated in the related condition. In phase one, items were presented randomly with the restraint that a given cue was always followed immediately by its respective target.

Phase two in this experiment was similar to the initial phase in Experiments 1 and 2. For 30 of the trials intact pairs were presented for 13 s and participants were instructed to study the cue–target pairs for a later cued recall test. For the other 30 trials a cue was presented with question marks appearing beside it. On these trials participants were told to use the word as a cue to recall a word from the first phase. They were told that the cue and the target had appeared in the earlier phase. To further encourage them to actually try to recall, participants in the unrelated condition were told that the target would have the same number of letters as the cue. In the related condition, participants were told that the target would be related or associated to the cue. Participants were encouraged to type in a response before 8 s had elapsed. The cue remained on the screen alone for 8 s, after which the cue and its paired target were presented for 5 s. At this point participants were to study the cue and target for a later cued-recall test. The instructions for this experiment were meant to encourage the participants to engage in an episodic retrieval mode in which they were seeking to retrieve the target from a spatiotemporal context (Tulving, 1983; Karpicke & Zaromb, 2010). As our interest was in unsuccessful retrieval attempts, the unintentional and brief encoding was used in the first phase. The participants across both conditions rarely successfully retrieved the correct target when presented with the cue. Only trials with unsuccessful retrieval attempts were included in the analysis. After completing the second phase, participants completed the 5 min country generation task. In the third phase, participants completed a cued-recall test. Each cue was randomly presented. Participants were told to type in the correct target that was paired with the cue in the second phase. Partici-



**Fig. 4.** Final cued-recall performance for Experiment 4. Error bars reflect SEM.

pants then pressed the “ENTER” key to begin the next trial. If they could not think of the target, they were allowed to type “XXX” and press the “ENTER” key to pass to the next trial.

### Results and discussion

A 2 (unrelated vs related)  $\times$  2 (read vs test) mixed model ANOVA revealed a significant main effect of relatedness,  $F(1, 28) = 5.92, p = .022, \eta_p^2 = .18$ , which was qualified by a significant interaction,  $F(1, 28) = 13.72, p = .001, \eta_p^2 = .33$ . The main effect of study was not significant,  $F(1, 28) = .37, p = .55$ . Follow-up comparisons revealed that the interaction was due to the read items being better remembered in the unrelated condition,  $t(14) = 2.43, p = .029, d = .31$ , whereas the test items were better remembered in the related condition,  $t(14) = -2.8, p = .014, d = .55$ , (Fig. 4). This pattern of results replicates the pattern of results found in Experiments 1A, 1B, and 2, and extends them by demonstrating that the benefit of unsuccessful retrieval is not limited to a semantic generation retrieval mode, rather it also benefits learning during episodic retrieval attempts. These findings suggest unsuccessful episodic retrieval attempts influence retention similarly to semantic generation. Nevertheless, one may contend that participants may have perceived the probability of recalling an item from phase one as very low and thus chose not to attempt episodic retrieval, and this may be particularly so for the unrelated condition where pairs only match on word length. We aimed to allay these concerns in the final experiment.

### Experiment 5

In this experiment we sought to further ensure that participants actually attempted retrieval by increasing the perceived likelihood of correctly retrieving targets for both related and unrelated pairs. We modified the three phase paradigm from Experiment 4 by including fewer cue–target pairs and presenting a subset of the cue–target pairs multiple times (i.e., 10 repetitions) during phase one; thus, in the read–test phase (phase two) participants should perceive correct retrieval as a probable event and would in turn be inclined to attempt episodic retrieval.

Further, we manipulated the relatedness of the pairs with-in subjects to place both unrelated and related pairs in a similar context where correct retrieval is probable.

### Method

#### Participants

Undergraduate students from the University of Georgia participated in exchange for partial credit toward a research requirement. Each person was either tested individually or in pairs at separate computers in sessions that lasted approximately 30 min. Thirty participants completed this experiment in which the association of the cue–target pairs was manipulated within subjects.

#### Materials

A subset of the materials from Experiment 4 were randomly selected and used to construct 48 cue–target pairs, half unrelated and half related. Additionally, half of both the unrelated and related pairs were made up of four letter words and the other half five letter words.

#### Procedure

This experiment employed three phases similar to Experiment 4 with a few modifications. All participants were presented with both unrelated and related pairs. In the initial phase, all words from the cue–target pairs were presented on the screen one at a time. Seventy-two words making up 36 of the cue–target pairs (18 unrelated and 18 related) were presented once for one second each. The 24 words from the other 12 pairs (6 unrelated and 6 related) were each presented 10 times (not sequentially) for one second each time. Repeated presentation of words were included to increase participants’ perceived likelihood of being able to recall words from this phase once they began the second phase that included a surprise test. There were 312 trials in phase one, and for each trial participants were instructed to read the word silently, and hence, they were unaware of a future memory test. All items were presented randomly with the restraint that a given cue was always followed immediately by its respective target. After completion of the first phase and as in Experiment 4, participants completed the STAI (Spielberger et al., 1970) and the TIPI (Gosling et al., 2003) which in total took about 5 min then the second phase began. The words that were presented in the first phase served as the cue–target pairs in the second phase. As in Experiment 4, the cue and target in each pair had the same number of letters and was either unassociated (unrelated pairs) or associated (related pairs).

Phase two in this experiment was similar to phase two in Experiment 4. For 24 of the trials intact pairs were presented for 13 s and participants were instructed to study the cue–target pairs for a later cued recall test. Half of these pairs were unrelated and half were related. Of the 12 pairs from both types nine of the cue and targets were presented once in phase one and three were presented 10 times in that initial phase. For the other 24 trials a cue was presented with question marks appearing beside it. On these trials participants were told to use the word as a cue to recall a word from the first phase. They were told that the cue and the target they were trying to recall had

appeared in the earlier phase and that the target had appeared immediately after the cue in that phase. To further guide their retrieval attempt, one of two prompts, *LETTERS* or *ASSOCIATE*, appeared above the cue to indicate whether the target had the same number of letters or was related, respectively. Participants were encouraged to type in a response before 8 s had elapsed. The cue remained on the screen alone for 8 s, after which the cue and its paired target were presented for 5 s. At this point participants were to study the cue and target for a later cued-recall test. Twelve of the tested pairs were unrelated and 12 were related. As was the case for the read trials, of the 12 pairs from both types nine of the cue and targets were presented once in phase one and three were presented 10 times in that phase. In addition to encouraging the participants to engage in an episodic retrieval mode (Tulving, 1983; Karpicke & Zaromb, 2010), this manipulation was meant to provide a situation in which participants actually felt that successful retrieval was possible so that they would attempt retrieval on the vast majority of test trials. That is, participants should have confidence that they can remember some of the targets, given that they experienced a number of them for 10 repetitions, and thus participants should be more inclined to attempt to retrieve the targets. As our interest was in unsuccessful retrieval attempts, the unintentional and brief encoding was used in the first phase. Accordingly, participants rarely successfully retrieved the correct target for both unrelated and related pairs that had only been presented once in phase one. Only trials with unsuccessful retrieval attempts were included in the analysis. After completing the second phase, participants completed the 5 min country generation task. In the third phase, participants completed a cued-recall test only for the pairs that had been presented once in the initial phase. Each of the 36 cues was randomly presented. Participants were told to type in the correct target that was paired with the cue in the second phase. Participants then pressed the “ENTER” key to begin the next trial. If they could not think of the target, they were allowed to type “XXX” and press the “ENTER” key to pass to the next trial.

### Results and discussion

A 2 (unrelated vs related)  $\times$  2 (read vs test) repeated measures ANOVA revealed a significant main effect of relatedness,  $F(1, 29) = 22.31$ ,  $p < .001$ ,  $\eta_p^2 = .44$ , which was qualified by a significant interaction,  $F(1, 29) = 13.87$ ,  $p = .001$ ,  $\eta_p^2 = .33$ . The main effect of study was not significant,  $F(1, 29) = .15$ ,  $p = .7$ . Follow-up comparisons revealed that the interaction was due to the read items being better remembered for the unrelated pairs,  $t(29) = 2.6$ ,  $p = .014$ ,  $d = .56$ , whereas the test items were better remembered for the related pairs,  $t(29) = -3.5$ ,  $p = .001$ ,  $d = .88$ , (Fig. 5). This pattern of results, which replicates the pattern observed in Experiments 1A, 1B, 2, and 4, extends the prior findings by demonstrating a similar influence of unsuccessful retrieval when each participant experienced both unrelated and related pairs and when participants were likely attempting episodic retrieval due to perceiving correct retrievals as probable. This notion that participants were indeed attempting episodic retrieval

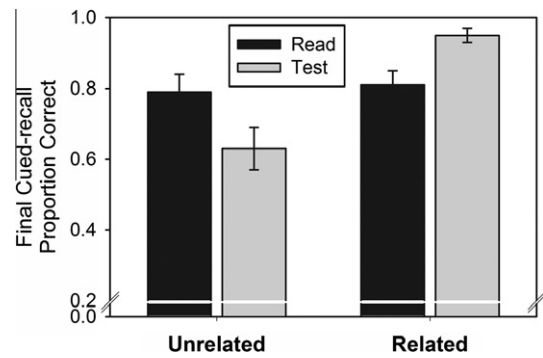


Fig. 5. Final cued-recall performance for Experiment 5. Error bars reflect SEM.

is supported by the finding that the highest percentage of correct retrievals across all experiments was found here. When considering all test items from the read-test phase, the correct target was retrieved for 1% and 15% of cues for the unrelated and related pairs, respectively. Moreover, even when retrieval was incorrect participants' recalls included other items from the phase one list on 26% and 9% of the trials for the unrelated and related pairs, respectively, providing further evidence that participants were attempting to episodically retrieve the target on test trials for both pair types. Thus, the influences of unsuccessful tests in episodic retrieval attempts followed the same patterns as was observed in semantic generation.

### False recall on final cued-recall tests across the experiments

Consideration of the intrusion errors output on the final cued-recall tests can help further elucidate the influence that unsuccessful retrieval attempts have on learning and retention. Unsuccessful tests may have a negative impact (e.g., increased intrusion errors) in addition to their detriment to learning when the pairs are unrelated. That is, although the impact of successful tests has typically been positive, negative impacts of multiple choice tests have recently been reported (Fazio et al., 2010; Roediger & Marsh, 2005). Fazio and colleagues found that incorrect answers selected on a multiple choice test persisted to subsequent cued-recall tests (immediate and delayed). Further, cued-recall intrusion errors increased linearly with the number of incorrect lures presented as alternatives on an initial multiple choice test, regardless of whether those lures were selected on the initial test. That is, a greater number of lures presented with multiple choice questions (e.g., 6 lures vs. 2 lures) resulted in more subsequent cued-recall intrusion errors for that material (see also McDermott, 2006). Thus, the presentation of additional potentially interfering information on tests can increase subsequent error rates. Here, the test pairs relative to the read pairs likely had more words active during study, as participants were generating a guess before being presented with the intact pair. As a result, the intrusion error rates could be increased for the tested pairs.

Recent evidence has demonstrated that, though intrusion errors occur rarely on memory tests, they are system-

**Table 2**  
Mean intrusions (SEM) output for Experiments 1, 2, 4, and 5.

	Intrusions					
	Extra-list		Intra-list		Generate	
	Read	Test	Read	Test	Read	Test
<i>Experiment 1A</i>						
Unrelated	.11 (.03)	.19 (.03)	.12 (.02)	.18 (.03)	–	.08 (.03)
Related	.24 (.04)	.12 (.02)	.03 (.01)	.02 (.01)	–	.04 (.01)
<i>Experiment 1B</i>						
Unrelated	.28 (.04)	.24 (.04)	.15 (.02)	.18 (.03)	–	.09 (.03)
Related	.33 (.04)	.12 (.02)	.05 (.01)	.02 (.01)	–	.06 (.01)
<i>Experiment 2</i>						
Unrelated	.27 (.03)	.21 (.03)	.11 (.02)	.21 (.02)	–	.06 (.02)
Related	.19 (.03)	.07 (.02)	.06 (.01)	.02 (.01)	–	.03 (.01)
<i>Experiment 4</i>						
Unrelated	.26 (.06)	.22 (.04)	.16 (.04)	.24 (.04)	–	.06 (.02)
Related	.25 (.06)	.09 (.03)	.06 (.02)	.03 (.01)	–	.06 (.03)
<i>Experiment 5</i>						
Unrelated	.08 (.03)	.18 (.04)	.13 (.03)	.14 (.04)	–	.06 (.02)
Related	.14 (.02)	.02 (.01)	.05 (.02)	.02 (.01)	–	.02 (.01)

Note. The – indicates the data are not available by design.

atic and informative of the mnemonic mechanisms involved (Unsworth & Brewer, 2010a, 2010b; Unsworth, Brewer, & Spillers, 2010). In the present study three types of intrusion errors could occur: extra-list, intra-list, and generate intrusions. Extra-list intrusions (ELIs) represent words output during cued-recall that were not presented on the study list. Intra-list intrusions (ILIs) represent words that were presented on the study list (either as cues or targets) but were output in response to a word that it was not paired with on the list. Generate intrusions (GIs) represent errors in which participants responded to a cue with the word that they initially generated or guessed rather than the correct target. A mixed model ANOVA with a within-subjects factor of study (read vs. test) and a between-subjects factor of relatedness (unrelated vs. related) was computed for both ELIs and ILIs separately for each experiment (repeated measures ANOVA was implemented in Experiment 5). Inferential statistics were not computed for GIs due to the unbalanced nature of these errors (i.e., GIs could only occur for test pairs).

Means for ELIs, ILIs, and GIs for Experiments 1A, 1B, 2, 4, and 5 are listed in Table 2. The ANOVA results for each experiment can be found in Appendix. Experiment 3 intrusions are not reported due to the absence of read pairs in that manipulation. Overall, the false recall analyses revealed informative findings about the influence of unsuccessful retrieval attempts that were followed by feedback. Across the experiments that included read versus test pair comparisons, participants in the related condition tended to output ELIs less often for test pairs than read pairs, whereas in the unrelated condition the difference was statistically nonexistent or in the opposite direction. This pattern was statistically confirmed by a significant interaction in each of those experiments except Experiment 2 (see Appendix). When mean ELIs were pooled across experiments (that implemented a between-subjects design) and analyzed using a mixed-model ANOVA with factors of study type and relatedness, significant main effects of

study type and relatedness were found,  $F(1, 218) = 53.33$ ,  $p < .001$ ,  $\eta_p^2 = .2$ ;  $F(1, 218) = 4.46$ ,  $p = .036$ ,  $\eta_p^2 = .02$ , respectively. Critically, a significant interaction was observed,  $F(1, 218) = 35.83$ ,  $p < .001$ ,  $\eta_p^2 = .14$ , which resulted because relative to read pairs the proportion of ELIs output by participants for test pairs was significantly lower in the related condition,  $t(109) = 10.14$ ,  $p < .001$ ,  $d = .91$ , and statistically equivalent in the unrelated condition,  $t(109) = .87$ ,  $p = .39$ . Though the Experiment 5 ELI means were not included in this analysis due to the within-subjects design implemented, the same critical interaction was observed in that experiment (see Supplementary material).

Unsworth and Brewer (2010a) found that intrusion errors were largely dependent on source monitoring processes (Johnson, Hashtroudi, & Lindsay, 1993) in which participants must accurately monitor the presented items at encoding and assess the qualitative characteristics of sampled items at retrieval to ensure they match characteristics from the list context and thus should be output. Here, unsuccessful tests on related pairs may have reduced ELIs by better delineating presented items from related information that was not presented, and thus aiding source monitoring processes. This observed influence of unsuccessful tests is consistent with recent findings of reduced intrusions (Szpunar, McDermott, & Roediger, 2008), enhanced recollection (Chan & McDermott, 2007), and enhanced source monitoring processing (Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010) for successfully tested items relative to untested items.

Identical analysis of pooled ILI means revealed similar results, with significant main effects of study type and relatedness being found,  $F(1, 218) = 6.2$ ,  $p = .013$ ,  $\eta_p^2 = .03$ ;  $F(1, 218) = 112.46$ ,  $p < .001$ ,  $\eta_p^2 = .34$ , respectively. A significant interaction was also found,  $F(1, 218) = 27.95$ ,  $p < .001$ ,  $\eta_p^2 = .11$ , that resulted because relative to read pairs ILIs occurred significantly less often for test pairs in the related condition,  $t(109) = 3.98$ ,  $p < .001$ ,  $d = .47$ , but



significantly more often for test pairs in the unrelated condition,  $t(109) = -4.15$ ,  $p < .001$ ,  $d = .46$ . Thus, for conditions in which unsuccessful tests are beneficial to correct recall (i.e., when the output error activates a conceptual network related to the target) unsuccessful retrieval attempts may improve the discriminability of the pairs on the list relative to just studying the pairs. The usefulness of the initial guess as a mediating cue in the related condition could serve to delimit the memory search set for a given cue, thus reducing the likelihood that other words from the list are in the search set and potentially sampled (e.g., Unsworth, 2008). Such information used to query memory in the unrelated condition may have been less diagnostic, resulting in an increased likelihood of sampling other items from the list; nevertheless, future work is needed to better understand how retrieval dynamics are influenced by prior unsuccessful retrieval attempts. Additionally, the ILI interaction was significant in just two of the experiments (see [Supplementary material](#)), so it awaits further replication and should be interpreted with caution. The consistent ILI finding across the experiments is that ILIs were lower in the related than the unrelated condition regardless of study type; that is, participants in the related condition were less likely to output other items on the list for the inappropriate cue. This effect may have resulted due to the fact that each target in the related condition was associated to its cue which could help participants on the final test to delimit their search to only words that were related to the presented cue. The lack of relatedness between each cue and target in the unrelated condition would decrease the distinction between studied items, resulting in an increased likelihood of outputting a presented list item for the inappropriate cue. Assessment of the GIs revealed that across experiments participants rarely output their generated guess on the final test (see [Table 2](#)), and the proportion of GIs did not substantially differ for the related and unrelated conditions, suggesting that initially generated errors were rarely confused with the correct target on the final test.

## General discussion

In these Experiments we demonstrated that unsuccessful retrieval attempts followed by feedback on tests can *both* enhance and impair learning. Research on retrieval practice have produced mixed results regarding the influence of unsuccessful retrieval attempts on memory, with reports of both detriments (Cunningham & Anderson, 1968) and benefits (Kornell et al., 2009; Richland et al., 2009). Accordingly, the mechanisms by which unsuccessful tests affect learning were uncharacterized. Here we demonstrated both impairments and enhancements and provided some reconciliation between these disparate findings by providing evidence for two mechanisms—elaborative semantic activation and errors as mediating cues—that are influential in the effect of unsuccessful retrieval on learning. Further, our data provide evidence against two other explanatory factors—error suppression and enhanced attention/arousal explanations. These results extend previous findings by demonstrating that unsuccessful tests enhance memory by reducing false

recall in addition to aiding retention and by revealing unsuccessful retrieval is influential in episodic retrieval attempts as well as semantic generation. Overall, these data provide the first elucidation of the explanatory mechanisms of unsuccessful tests and reveal important boundary conditions for its benefits.

In Experiment 1, when learning related (weak-associate) word pairs, unsuccessful retrieval attempts led to greater retention than did spending an equal amount of time studying the intact pairs. However, when learning unrelated word pairs, unsuccessful retrieval attempts resulted in worse later memory performance than did studying the pairs for the same amount of overall time (Exp. 1A; 13 s) or for the same amount of time that feedback lasted after unsuccessful tests (Exp. 1B; 5 s). The same pattern was observed when the instructions for producing an unsuccessful retrieval were switched between the unrelated (i.e., guess a word that is associated) and related (i.e., guess a word that has the same number of letters) conditions, revealing that the level of processing of the cue cannot explain the influence of unsuccessful tests on learning. In Experiment 3, on the final test participants were able to recall their initial guess a majority of the time and their guess elicited the target more readily for the related pairs. Thus, initial errors seem to have not been suppressed, but rather the incorrect guesses appear to have served as mediating cues for retrieving the target (primarily for the related pairs). Experiments 4 and 5 revealed the same influence of unsuccessful retrieval attempts when participants had episodic experience with the words and when correct retrievals were probable and one could be fairly certain that participants attempted episodic retrieval, thus extending the influence of unsuccessful tests to an episodic retrieval mode. Experiment 5 further extended these results by demonstrating the same pattern of findings when participants encountered both unrelated and related pairs.

Our finding of deleterious effects of unsuccessful retrieval attempts with unrelated word pairs is similar to the impairments reported by Cunningham and Anderson (1968) using number-trigram pairs, in that there is not a straightforward semantic relationship between the cue and target in either study. Thus, in both studies not only were participants unable to produce the correct target when initially presented with the cue, the guess they did produce was also unlikely to be conceptually related to the target. The unrelated, erroneous guess may in turn produce interference (Postman & Underwood, 1973), and in such situations minimization of errors during learning (Clare & Jones, 2008) may be optimal. Unsuccessful retrievals in which a conceptual or semantic network that is common to the question and answer is unable to be activated may be practically akin to students seeking to answer questions in a subject area or knowledge domain in which they have no experience or prior knowledge (e.g., an introductory college course). According to the current data, administering pre-tests (i.e., testing prior to material exposure; Richland et al., 2009) in a novel domain in which one does not possess enough prior relevant knowledge to produce educated guesses that activate a conceptual network relevant to the correct answer may do more harm than good. Comparisons of the influences of tests that

would be largely unsuccessful (e.g., pre-tests) with educationally relevant material at the outset of an introductory versus an upper level class (i.e., students should have prior knowledge of the subject matter) on a given topic would provide a better examination of this prediction.

The observations of enhancements for unsuccessfully tested relative to studied related cue–target pairs, and detriments for tested relative to studied unrelated pairs, lends clear support to the notion that the benefit of unsuccessful tests is reliant on the elaborative activation of a related semantic network which through spreading activation (Anderson, 1983) may provide a fertile context for encoding the target (Kornell et al., 2009). The notion that semantic knowledge plays a role in the benefit of unsuccessful retrieval attempts to learning is consistent with other demonstrations of the dependency of episodic encoding on prior knowledge that is congruent with the to-be-learned information (DeWitt, Knight, Hicks, & Ball, *in press*; Kan, Alexander, & Verfaillie, 2008). Results from Experiment 2 provide a further delineation of this elaborative processing mechanism by revealing that the level of processing of the cue is not the primary contributing factor. That is, even when participants in the unrelated condition were guessing a word related to the cue (and thus processing the semantic meaning of the cue) retention was lower for the test relative to the read pairs. The elaborative semantic activation encouraged by unsuccessful retrieval prior to and during encoding may indeed be reactivated during final test retrieval (i.e., cued-recall test) and serve to facilitate access to the target (e.g., Meade, Watson, Balota, & Roediger, 2007). Similarly, the current study demonstrates that unsuccessful retrieval attempts followed by feedback partly facilitate retention by establishing a mediating cue(s) that provides a link between the cue and target and aids memory search for the target when presented with the cue (Kornell et al., 2009; Pyc & Rawson, 2010; Soraci et al., 1999). The benefit to target recall as a function of recall of one's initial guess was only present for the related pairs, suggesting that the semantic activation and mediating cue accounts may be dependent upon one another. Further research is needed to determine the extent to which these explanatory mechanisms are dissociable and which mechanism accounts for more variance in the benefit of unsuccessful tests to retention.

Interestingly, the explanatory accounts of unsuccessful retrieval consistent with the present data are similar to recent theoretical proposals of the mechanisms thought to underlie the benefit of *successful* retrievals to retention. In a study by Carpenter (2009), participants were tested on or studied previously seen strong and weak cue–target pairs, and she found that the weak tested pairs, which would theoretically elicit the most elaborate activation of the related semantic network, elicited better retention on a final test. Thus, Carpenter (2009) proposed that the elaborative activation of information may underlie the testing effect. Consistent with this line of reasoning, Pyc and Rawson (2010) have found that successful tests facilitate retention, in part, by enhancing the effectiveness of mediating information between the cue (question) and target (answer; see “Introduction”), and Carpenter (2011) recently demonstrated the highly semantic nature of such

mediators in the testing effect. The striking similarity between these findings for successful tests and the results reported here for unsuccessful tests suggests that the facilitation to retention through successful and unsuccessful retrievals may be enacted by a common mechanism. Thus, the benefits of testing in general may be more dependent on the process of attempting to retrieve the answer rather than the actual acquisition of the correct answer; nevertheless, future research is needed to examine if unsuccessful retrievals can be as facilitative as successful retrievals.

Two other potential mechanisms have been proposed to explain the benefits of unsuccessful retrieval, neither of which was supported by our results. First, it has been suggested that incorrect guesses may suppress erroneous retrieval routes (Carrier & Pashler, 1992). This view would predict that participants' generated guesses should be largely inaccessible during the final test and that correct target recall should be inversely related to guess retrieval. However, generated guesses were retrieved on the majority of the trials and this occurred for both related and unrelated pairs. Furthermore, in conditions when unsuccessful retrieval was beneficial (i.e., with related pairs), correct target recall was associated with successful guess retrieval. Thus, unsuccessful retrieval of related information may actually strengthen correct retrieval routes, through cue mediation, rather than suppress erroneous retrieval routes. Also, it has been proposed that initial testing may encourage participants to enhance attentional processing during the learning phase (e.g., Chan & Langley, 2010), thus unsuccessful retrieval attempts may result in increased attention to encoding of the correct cue–target pair. However, unsuccessful testing only benefitted performance for related but not unrelated cue–target pairs, suggesting that increased attention is not the primary contributor to the benefit of unsuccessful retrieval. Rather, the facilitative effects of unsuccessful retrieval appear to be due to activation of related information within a semantic network during encoding (and possibly retrieval), as well as enhanced effectiveness of mediating cues during retrieval.

A demonstration of the benefits from unsuccessful retrieval attempts on subsequent recall performance following delayed feedback would be of high educational relevance. In the current study, participants first unsuccessfully output a target item and then were immediately presented with the correct cue–target pair. However, in the classroom setting it is not often during testing that immediate feedback is given. Currently, our results are limited to when feedback is immediate, for example, when a student answers a question incorrectly in class and the professor immediately corrects the student. Perhaps an important endeavor for future research in elucidating the mechanisms of unsuccessful retrieval would be to examine whether the benefits (and detriments) persist when feedback is delayed (akin to taking a test and then reviewing it at the end of class). To the extent that elaborative semantic activation (which is typically short-lived; e.g., Neely, 1977) at encoding is a primary contributing factor of unsuccessful retrieval attempts, then benefits may not be observed for delayed feedback following unsuccessful retrieval. However, previous research on the testing effect (i.e. successful retrieval) suggests that both immediate

and delayed feedback increase correct and decrease incorrect answers in multiple choice tests (e.g. Butler & Roediger, 2008) and in cued-recall (Pashler, Cepeda, Wixted, & Rohrer, 2005). Thus, perhaps an important direction in understanding the benefits of unsuccessful retrieval would be to compare immediate and delayed feedback on subsequent test performance. Likewise, comparing both immediate and delayed feedback to no feedback may highlight important aspects of learning and testing in the classroom setting.

## Conclusion

Most educators' prominent objection to increasing the number of tests in the classroom is the concern that errors output on tests will have a negative impact and lead to future errors (Pashler et al., 2007). Here we provide further evidence that unsuccessful tests in which errors are retrieved can enhance retention; however, we also demonstrate a key boundary condition, in that unsuccessful tests appear to be harmful to learning when the guessed information is not related to the answer and does not provide a meaningful link between the question and the answer. In general, the current findings suggest the influence of tests—whether successful or unsuccessful—may be supported by common or similar cognitive mechanisms.

## Acknowledgments

This research was supported by the John and Mary Franklin Foundation through a fellowship awarded to Justin B. Knight. We would like to thank Henry Roediger and two anonymous reviewers for their thoughtful comments on a previous version of the paper. We also appreciate the diligent efforts of Kristine Hughes, Adetola Adeyeye, Ross Beall, Kevin Davis, Renu Gulve, Janelle Burns, Anastasia Bobilev, Kristin Otis, Bethany Martin, and Heather Paine in assisting with data collection.

## A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jml.2011.12.008](https://doi.org/10.1016/j.jml.2011.12.008).

## References

- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3), 261–295.
- Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyala Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.
- Brewer, G., Marsh, R., Meeks, J., Clark-Foos, A., & Hicks, J. (2010). The effects of free recall testing on subsequent source memory. *Memory*, 18(4), 385–393. doi:10.1080/09658211003702163.
- Butler, D. C., & Peterson, D. E. (1965). Learning during "extinction" with paired associates. *Journal of Verbal Learning and Verbal Behavior*, 4(2), 103–106.
- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6), 1563–1569.
- Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(6), 1547–1552. doi:10.1037/a0024140.
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19(5), 619–636. doi:10.1002/acp.1101.
- Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, 23(6), 760–771. doi:10.1002/acp.1507.
- Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, 13(5), 826–830.
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633–642.
- Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, 18(1), 49–57. doi:10.1080/09658210903405737.
- Chan, J. C. K., & Langley, M. M. (2010). Paradoxical effects of testing: Retrieval enhances both accurate recall and suggestibility in eyewitnesses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/a0021204.
- Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2), 431–437. doi:10.1037/0278-7393.33.2.431.
- Clare, L., & Jones, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology Review*, 18(1), 1–23. doi:10.1007/s11065-008-9051-4.
- Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, 125(2), 159–180.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684.
- Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104(3), 268–294.
- Cunningham, D. J., & Anderson, R. C. (1968). Effects of practice time within prompting and confirmation presentation procedures on paired associate learning. *Journal of Verbal Learning and Verbal Behavior*, 7, 613–616.
- DeWitt, M. R., Knight, J. B., Hicks, J. L., & Ball, B. H. (in press). The effects of prior knowledge on the encoding of episodic contextual details. *Psychonomic Bulletin & Review*.
- Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, 38(4), 407–418. doi:10.3758/mc.38.4.407.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. Jr., (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504–528. doi:10.1016/s0092-6566(03)00046-1.
- Guthrie, E. R. (1952). *The Psychology of Learning* (Rev. ed.). New York: Harper.
- Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology*, 75(2), 194–209.
- Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, 83(2), 340–344.
- Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6), 1441–1451. doi:10.1037/a0020636.
- Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114(1), 3–28.
- Kan, I. P., Alexander, M. P., & Verfaillie, M. (2008). Contribution of prior semantic knowledge to new episodic learning in amnesia. *Journal of Cognitive Neuroscience*, 21(5), 938–944.
- Karpicke, J. D., & Roediger, H. L. III., (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968. doi:10.1126/science.1152408.
- Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, 62(3), 227–239. doi:10.1016/j.jml.2009.11.010.
- Kornell, N., Hays, J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 989–998.
- Marsh, E. J., Agarwal, P. K., & Roediger, H. L. III., (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, 15(1), 1–11. doi:10.1037/a0014721.

- Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, 14(2), 194–199.
- McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland, D. E. Rumelhart, T. P. R. Group (Eds.), *Parallel distributed processing: Explorations in microstructure of cognition. Psychological and biological models* (Vol. 2, pp. 170–215). Cambridge, MA: MIT.
- McDaniel, M. A., Agarwal, P. K., Huelsner, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103(2), 399–414. doi:10.1037/a0021782.
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19(4), 494–513. doi:10.1080/09541440701326154.
- McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(2), 371–385.
- McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, 14(2), 200–206.
- McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, 34(2), 261–267.
- Meade, M. L., Watson, J. M., Balota, D. A., & Roediger, H. L. III, (2007). The roles of spreading activation and retrieval mode in producing false recognition in the DRM paradigm. *Journal of Memory and Language*, 56(3), 305–320. doi:10.1016/j.jml.2006.07.007.
- Mueller, M. M., Palkovic, C. M., & Maynard, C. S. (2007). Errorless learning: Review and practical application for teaching children with pervasive developmental disorders. *Psychology in the Schools*, 44(7), 691–700. doi:10.1002/pits.20258.
- Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibition less spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106(3), 226–254.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme and word fragment norms. [web.usf.edu/FreeAssociation](http://web.usf.edu/FreeAssociation).
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 3–8. doi:10.1037/0278-7393.31.1.3.
- Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., McDaniel, M. A., & Metcalfe, J. (2007). Organizing instruction and study to improve student learning (NCER 2007–2004).
- Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, 14(2), 187–193.
- Postman, L., & Underwood, B. L. (1973). Critical issues in interference theory. *Memory & Cognition*, 1, 19–40.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4), 437–447. doi:10.1016/j.jml.2009.01.004.
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330, 335.
- 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), 163–178.
- Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, 15(3), 243–257. doi:10.1037/a0016496.
- Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. doi:10.1016/j.tics.2010.09.003.
- Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210.
- Roediger, H. L., III, & Marsh, E. J. (2005). The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155–1159. doi:10.1037/0278-7393.31.5.1155.
- Roediger, H. L., III, & Payne, D. G. (1982). Hypernesia: The role of repeated testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(1), 66–72.
- Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, 39(5), 406–412.
- Skinner, B. F. (1958). Teaching machines. *Science*, 128(3330), 969–977.
- Soraci, S. A., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory and Language*, 41, 541–559.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *The STAI manual for the State-Trait Anxiety inventory*. Palo Alto, California: Consulting Psychologists 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(6), 1392–1399. doi:10.1037/a0013082.
- Terrace, H. S. (1963). Errorless transfer of a discrimination across two continua. *Journal of the Experimental Analysis of Behavior*, 6(2), 223–232.
- Tulving, E. (1983). *Elements of Episodic Memory*. Oxford: Clarendon.
- Unsworth, N. (2008). Exploring the retrieval dynamics of delayed and final free recall: Further evidence for temporal-contextual search. *Journal of Memory and Language*, 59(2), 223–236. doi:10.1016/j.jml.2008.04.002.
- Unsworth, N., & Brewer, G. A. (2010a). Individual differences in false recall: A latent variable analysis. *Journal of Memory and Language*, 62(1), 19–34. doi:10.1016/j.jml.2009.08.002.
- Unsworth, N., & Brewer, G. A. (2010b). Variation in working memory capacity and intrusions: Differences in generation or editing? *The European Journal of Cognitive Psychology*, 22, 990–1000. doi:10.1080/EJCP.22.6.990.
- Unsworth, N., Brewer, G. A., & Spillers, G. J. (2010). Understanding the dynamics of correct and error responses in free recall: Evidence from externalized free recall. *Memory & Cognition*, 38(4), 419–430. doi:10.3758/mc.38.4.419.
- Verde, M. F. (2009). The list-strength effect in recall: Relative-strength competition and retrieval inhibition may both contribute to forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1), 205–220. doi:10.1037/a0014275.