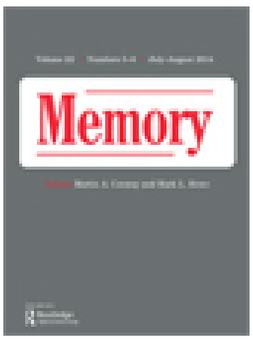

Memory

Routledge
Taylor & Francis Group

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

Retrieval attempts enhance learning regardless of time spent trying to retrieve

Kalif E. Vaughn, Hannah Hausman & Nate Kornell

To cite this article: Kalif E. Vaughn, Hannah Hausman & Nate Kornell (2016): Retrieval attempts enhance learning regardless of time spent trying to retrieve, Memory, DOI: 10.1080/09658211.2016.1170152

To link to this article: http://dx.doi.org/10.1080/09658211.2016.1170152

MEMORY, 2016
http://dx.doi.org/10.1080/09658211.2016.1170152

Routledge
Taylor & Francis Group

Retrieval attempts enhance learning regardless of time spent trying to retrieve

Kalif E. Vaughn[a], Hannah Hausman[b] and Nate Kornell[a]

[a]Department of Psychology, Williams College, Williamstown, MA, USA; [b]Department of Psychology, Colorado State University, Fort Collins, CO, USA

ABSTRACT

Attempting to retrieve information from memory is an engaging cognitive activity. We predicted that people would learn more when they had spent more time attempting to retrieve. In experiments 1a and 1b, participants were shown trivia questions for 0, 5, 10, or 30 seconds and then the answer was revealed. They took a final test immediately or after 48 hours. Retrieval enhanced learning, but the length of the retrieval attempt had no effect (i.e., final test performance was equivalent in the 5-, 10-, and 30-second conditions and worse in the 0-second condition). During the initial retrieval attempt, more time did increase recall, suggesting that participants continued to engage in productive retrieval activities when given more time. Showing the answer for longer (7 versus 2 seconds) increased learning in Experiments 2a and 2b. Experiment 3 examined the effect of retrieval success and Experiment 4 replicated the results using different materials. These results have direct implications for current theories of retrieval.

ARTICLE HISTORY
Received 25 October 2015
Accepted 18 March 2016

KEYWORDS
Retrieval effort; retrieval success; testing; memory

According to the Billboard charts, what group claimed both of the top two spots on the list of the best-selling albums in the US in 1967? If you want to learn the answer to this question, a large literature shows that attempting to retrieve the answer from memory (i.e., retrieval practice) is more effective than simply being told the answer (for reviews, see Roediger & Butler, 2011; Roediger & Karpicke, 2006). But an unexplored question is, how long should you think about it? You could spend a lot of time because there is a lot to think about—new albums were released by Bob Dylan, The Doors, Pink Floyd, The Rolling Stones, and the Who, and multiple albums were released by Aretha Franklin, The Jimi Hendrix Experience and Jefferson Airplane, not to mention the Beatles, who released Sgt. Peppers and Magical Mystery Tour. On the other hand, if your goal is to learn the answer with minimal fuss, it would be more efficient to avoid spending too much time trying to think of the answer and just find out what it is.[1]

The research we report examined whether learning is affected by how long you spend thinking about a question before you find out the answer. Because a retrieval attempt involves active cognitive processing of relevant information, we predicted that spending more time attempting to retrieve might lead to more learning. For example, you might remember the answer better a few days later if you had spent 30 seconds trying to think of it than if you had spent only 5 seconds. In the present experiments, we manipulated the length of the retrieval attempts during the initial test trials, such that some items received more versus less retrieval time.

Reasons why more retrieval time might improve learning

Increased retrieval effort

A longer time spent in retrieval mode (e.g., Tulving, 1983) might increase retrieval effort. For example, attempting to retrieve for 5 seconds would, presumably, involve less effort than attempting to retrieve with equal effort for 5 seconds and then continuing the attempt for another 5 seconds. According to the *retrieval effort hypothesis*, "difficult but successful retrievals are better for memory than easier successful retrievals" (Pyc & Rawson, 2009, p. 473). This claim is specifically about the difficulty of successfully retrieving the memory, but if we extend it to time spent trying to retrieve, then the retrieval effort hypothesis makes two relevant predictions: (1) retrieval attempts will enhance learning, and (2) the attempts will enhance learning to a greater extent when learners exert more effort. In short, REH predicts that more time in retrieval mode should enhance learning.

A few researchers have examined the extent to which time spent attempting to retrieve influences subsequent performance. For instance, Auble and Franks (1978) had participants read sentences that needed a cue word in order to be understood (e.g., "The party was stalled because the wire straightened", is meaningless without the cue word, "corkscrew"). Auble and Franks manipulated

CONTACT Kalif E. Vaughn kev1@williams.edu Bronfman Science Center, Williams College, Williamstown, MA 01267, USA

when the cue was presented, with the possibilities being that the cue was presented 5 seconds after the sentence, 5 seconds before the sentence, or embedded within the sentence itself. They found that embedding the cue within the sentence resulted in worse memory performance than either presenting the cue before or after the sentence was read (with no difference between those two conditions; see Experiment 2). They replicated this finding across experiments, and in Experiment 4 once again showed that delaying the presentation of the cue for 5 seconds was associated with better performance than presenting the cue without a delay (which minimises the chance of spontaneous retrieval attempts). They concluded that, when the cue was not embedded within the sentence and was shown after a delay, the sentence required more effort to understand, and that this additional "effort toward comprehension" boosted subsequent recall. These results seem to support REH.

Although the results from Auble and Franks (1978) suggest that retrieval effort and additional time spent trying to retrieve the correct answer should enhance subsequent recall performance, there are a few key differences between their study and the current one. First, Auble and Franks (1978) did not manipulate time spent in retrieval mode to the same extent as we did. Auble and Franks either showed the answer immediately (a study control condition) or with a delay of 5 seconds (either before or after the sentence was presented). Comparing the 5-second condition to the study control condition (0-second condition) is a way to tell whether retrieval is better than no retrieval, but does not answer the question of whether more time spent in retrieval mode is better than less time spent in retrieval mode. In the current experiments, to examine the benefits of prolonged retrieval, we had participants attempt retrieval for 0, 5, 10, or 30 seconds.

Related to Auble and Franks (1978), Gardiner, Craik, and Bleasdale (1973) had experimenters read out the definition of a word and timed how long it took the participant to recall the correct target. Of interest, they measured how long it took participants to recall the correct answer during practice (with a maximum of 1 minute to retrieve each answer). They found that items recalled between 15 and 60 seconds were always recalled better on the final test as compared to items recalled in less than 15 seconds. This finding is surprising, especially given the extent to which item selection effects may have been apparent in the study (i.e., easy items presumably take less time to recall than difficult items, which would suggest that items recalled right away should be better remembered on a final test because they are easier items).

The results from Gardiner et al. (1973) were conceptually replicated by Benjamin, Bjork, and Schwartz (1998). Benjamin et al. presented trivia questions to participants and timed how long it took them to retrieve the correct answer. Participants then predicted how well they would freely recall the answer on a subsequent test. In general, the more quickly participants were able to retrieve the correct answer (i.e., the greater the retrieval fluency), the better they thought they would freely recall the answer on a later test. However, the opposite finding was observed: participants were better able to freely recall the answer when it took longer to retrieve the correct answer on the initial test. Thus, more time searching memory for the correct answer was associated with better free recall performance.

As with Auble and Franks (1978), the results from Gardiner et al. (1973) and Benjamin et al. (1998) provide evidence that more time in retrieval mode improves learning, and provide support for the REH. However, note that these studies assessed memory via free recall tests (e.g., recall as many of the words in the sentence as you can remember), whereas we always used cued recall tests in the present experiments (e.g., by presenting the same trivia question again). Additionally, within these experiments, time in retrieval mode was assessed but not manipulated. The potential concern is that participants spent longer amounts of time on items that required longer amounts of time (i.e., on the items they found more difficult). As such, the results could be subject to item difficulty effects. In the current experiments, we always randomly assigned the amount of retrieval time a particular item received, minimising item difficulty effects. Given these important differences, it remains an open question whether the same pattern of results will emerge in the current experiments.

One could argue that more time in retrieval mode might lead to less effort, not more. For example, a subject who knows she has only 5 seconds might try very hard, whereas someone who knows she has 30 seconds might be less focused. We designed our procedure with this possibility in mind: in the current experiments, participants were not told how long they would have to retrieve on any given trial (out of a possible 5, 10, or 30 seconds). As such, in the retrieval conditions, the first 5 seconds of each trial type were functionally the same and required the same amount of effort. The key difference is that participants either stopped trying to retrieve after those initial 5 seconds, or kept trying to retrieve the answer for an additional 5 or 25 seconds. In other words, because participants did not know how long they would have to retrieve, the degree of effort per second did not differ across conditions, but the amount of time spent trying to retrieve did. Thus, it was not like telling one group they had to run one lap and another that they had to run six (in which case the six-lap runners would run slower); it is like telling everyone to run a lap and then telling a randomly selected subset of runners that they have to run another five laps.

Desirable difficulties

There is an overwhelming amount of evidence that difficulty and challenge during learning tend to enhance learning in a variety of situations (e.g., Bjork, 1994; Bjork & Bjork, 2011). For instance, Rohrer and Taylor (2007) had

participants learn math formulas used to calculate the volume of various shapes. In the blocking condition, participants focused exclusively on learning how to calculate the volume for one shape at a time before moving on to the next shape. In the interleaving condition, participants practised learning how to calculate volume for shapes in a randomised order. The key difference between the conditions is that, in the blocking condition, mastery is achieved quickly because participants always know which formula to apply for a particular practice problem. In the interleaving condition, achieving mastery is much more difficult because participants always received a random practice problem, and thus never knew which formula they would need to use until the problem was shown. On a final test, participants were much better at calculating the volume of shapes when they had used an interleaving versus blocking schedule during initial learning.

The benefits of interleaving have been demonstrated with other types of material as well. Kornell and Bjork (2008) had participants learn painting styles for 12 artists based on a representative sample of 6 paintings. During practice, half of the artist's paintings were presented consecutively, one after the other, in a blocked run of trials. The other half of the paintings were presented in an interleaved fashion, such that one artist's painting was followed by a separate artist's painting, and so on. Participants thought that they would perform much better on a final test (which required matching up a novel painting to the correct artist) when they studied the paintings in a blocked order, but the interleaving order was far superior.

In addition to interleaving, the desirable difficulty pattern has also been shown with other types of manipulations. For instance, Smith, Glenberg, and Bjork (1978) manipulated the type of context in which the information was originally encoded (either listening to an audio cassette or studying the words visually on a projector). Additionally, the physical locations of these manipulations also varied (i.e., the audio cassette was played in a separate room from the projector slides, and these locations differed markedly in terms of appearance and style). All participants were exposed to the information a second time, and the second context either matched the first or differed from the first. On a surprise free recall test 3 hours later in a neutral context, participants recalled more words when they had been exposed to both contexts during learning (and not just one context; see their Experiment 1).

Desirable difficulty patterns have also emerged with respect to children's skill acquisition. Kerr and Booth (1978) had children practice throwing beanbags at a target on the floor. Some of the children practised at a fixed-distance away from the target (e.g., 3 feet), whereas other children practised at varying distances during practice. The varied practice group performed better on the final test, even though the final test distance always matched the distance used in the fixed-distance group (e.g., 3 feet). This finding suggests that the more difficult practice condition (i.e., the varied practice condition) allowed for more error correction to occur after each throw, which enhanced final test performance.

If the driving force behind the desirable difficulty pattern is that desirable difficulties cause people to engage in more active processing during initial encoding and/or retrieval, then we may see that longer search times (which would presumably involve active processing that lasts for more time) confer a delayed memory benefit compared to shorter search times. To summarise, the desirable difficulty framework and retrieval effort hypothesis seem consistent with the prediction that more time in retrieval mode should lead to more learning.

Reasons why retrieval time might not matter (or could hurt performance)

Spontaneous retrieval

There is an empirical reason to predict that time in retrieval mode might not matter. Benassi, Overson, and Hakala (2014) reported a series of studies with middle-school students in which retrieval practice was not more effective than simply reading the correct answer. Participants were briefly shown key terms on the screen (e.g., "opaque" was presented for 1 second in Experiments 1 and 2), followed by either a revealing of the correct definition (e.g., "not letting light through") or the appearance of an answer box with instructions to type in the correct definition. It is possible that the presentation condition was unintentionally converted into a retrieval condition because of the very brief time that the question was shown alone (e.g., seeing "opaque" alone on the screen may automatically trigger a search in memory for the correct definition), thus negating the difference between the presentation and retrieval conditions. Consistent with this idea, Benassi et al. (2014) did find a retrieval benefit when, during a presentation trial, they showed the question and answer simultaneously using word pairs (thereby minimising their ability to practice retrieval during the presentation trials). It might seem unlikely that participants could think of the answer (or even read the whole question on some trials) before the answer was shown; nonetheless, they might have read the question in retrieval mode because they knew the answer was not visible yet. There is evidence that participants can benefit from spontaneously generating answers from memory even when experimental conditions do not demand it (e.g., DeWinstanley & Bjork, 2004).

Irrelevant semantic activation

A final possibility is that increasing the amount of time in retrieval mode could hurt final test performance if learners activate the wrong information during the retrieval attempt. For example, if a learner were asked "what is the largest type of bear on Earth", she might conjure up a grizzly bear and spend time thinking of its features (it is brown, lives among trees, and so forth). Extra time spent in retrieval mode could interfere with her ability to learn

the correct answer—polar bear—when it mismatches the features she has been thinking about.

There are two reasons to suspect that additional time in retrieval mode will not hurt final performance, even if the wrong information is activated during the retrieval attempt. First, studies have demonstrated that failing to retrieve the correct answer can actually enhance memory, even without a prior study phase (e.g., Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, Hays, & Bjork, 2009; Vaughn & Rawson, 2012). Second, research has demonstrated that high-confidence errors are easily corrected (e.g., Butterfield & Metcalfe, 2001), such that in the example above, more certainty that the answer is "grizzly bear" might actually enhance learning of polar bear.

The present experiments

In the present experiments participants completed a study phase followed by a test phase. During the study phase, participants were either shown a trivia fact or they were initially shown the trivia fact with the answer omitted (see below) and given a fixed amount of time (5, 10, or 30 seconds) to try to retrieve the answer, after which they were shown the correct answer. On the final test, which occurred either immediately or after a two-day delay, they were asked the same questions again. This same basic procedure was used in each experiment, although new materials were used in the final experiment.

Terminating search early

Our experimental logic assumes that participants will be engaged in a retrieval attempt for a manipulated length of time on each test trial; however, participants might terminate their search quickly (e.g., after a few seconds) if they promptly think of an answer they are sure is correct (e.g., "One of Shakespeare's plays is called Romeo and _____"), or if they are so stumped that they give up right away (e.g., "Theodore Kerabatsos's favorite sport was _____"). If our participants did terminate their search early (e.g., if no one ever tried for more than 5 seconds), it would render all of the retrieval conditions equivalent in terms of their procedures, which would seriously undermine our studies. However, we were able to test for early termination by examining initial performance; early termination predicts that participants should not do better on the initial test when they are given more time to retrieve (to foreshadow, they did better). Furthermore, Experiment 4 was specifically designed to test the early termination hypothesis.

The two-stage framework

The process of learning from retrieval can be divided into two stages: (1) the retrieval attempt prior to the answer becoming available, and (2) processing of the answer (Kornell, Klein, & Rawson, 2015). Some readers might prefer to restrict the use of the term "retrieval process" to Stage 1, and refer to Stage 2 using another term (such as post-retrieval processing). To be clear, the two-stage framework refers to the process of *learning* from retrieval, not only the retrieval attempt itself. There is clear evidence that learning occurs during Stage 1 (e.g., Karpicke & Roediger, 2008) and Stage 2 (e.g., Arnold & McDermott, 2013). We are dividing the retrieval process into these two stages because learning occurs during both of them, and as will become evident later in the manuscript, we are interested in how the amount of time in either stage affects learning. But for now, our primary question is whether the amount of time spent in Stage 1 influences learning. In all of the present experiments, we manipulated the length of time participants spent in retrieval mode during study (0, 5, 10, or 30 seconds, where 0 seconds represents the presentation condition). We also asked whether the amount of time spent in Stage 2 matters, in Experiments 2a and 2b, by manipulating the length of time participants were shown the correct answer.

Experiments 1a and 1b

Experiments 1a and 1b used the basic procedure described above. Final test performance was measured immediately (Experiment 1a) or two days later (Experiment 1b). We used trivia questions as learning materials for two reasons. First, trivia questions are broad enough in nature to evoke retrieval from semantic memory (i.e., participants reading the question will try to sift through their general knowledge database as they search for the solution). For instance, we asked "The car that could travel through time in the film 'Back to the Future' was a _____." Compared to a word pair like frog-pond, this question should activate a richer set of semantic information, including memories of the film, cars from the 1980s, and information about the Hollywood milieu in 1985. Second, trivia questions help combat the aforementioned problem that participants may exhaust their semantic activation relatively quickly. By tapping into general knowledge on various broad topics (e.g., Hollywood), we are providing participants with the best chance at continually and productively activating semantic information during a retrieval attempt that might last up to 30 seconds.

Notice that with trivia questions, we do not need to provide an initial study phase for the information. Rather, given that the questions are general knowledge questions, we can start each experiment with a test trial (followed by eventual feedback, as described above). Other experimenters have used this "pretesting paradigm" (e.g., Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, Hays, & Bjork, 2009; Vaughn & Rawson, 2012), but typically it has been used to ensure that participants fail on the initial test trial and then to examine whether unsuccessful retrieval attempts benefit learning compared to restudy. Our purpose in using the

pretesting paradigm was to maximise the amount of effort that participants exert during a retrieval attempt (which will be higher without a prior study phase). We return to this issue within the "General Discussion".

Experiment 1a

Method

Participants

Forty-seven participants (30 female, 17 male; median age = 33 years, range = 19–66 years) completed the experiment on Amazon's Mechanical Turk. Participants were compensated $4.00 for completing the experiment. All participants reported being fluent English speakers living in the US (except for one participant who did not answer the fluency question). We included this participant in the analyses.[2]

Materials

Sixty trivia facts were used. The trivia facts reflected general knowledge across a variety of topics (e.g., mythological creatures, sports, geography). Example facts include: "Dominos were invented in the country of China", and "Hemlock is the kind of poison that Socrates took at his execution" (see Appendix A for a full list of the trivia facts used). During test trials, the trivia fact was converted to a fill-in-the-blank question (e.g., "Dominos were invented in the country of _____", and "_____ is the kind of poison that Socrates took at his execution").

Procedure

The experiment was conducted online. The initial phase of the experiment consisted of test trials. The 60 trivia facts were divided equally among four possible test trial timings: 0, 5, 10, and 30 seconds (i.e., 15 trivia facts were randomly assigned to each timing condition). The order of questions and the order of question timing were both random.

In the 5-, 10-, and 30-second conditions, participants were presented with the fill-in-the-blank question and a text box so that they could type in their answers. Participants could type their answer at any point until the test trial ended. The participants did not find out how long they had to answer until the test trial ended, which was important because it meant that the first 5 seconds of all trials were the same, from the participants' perspective, and the first 10 seconds of the 10- and 30-second conditions were the same. This meant that the amount of effort participants put into retrieval had to be at least as great in the longer conditions as the shorter ones, and presumably it was greater because of the additional time.

Once the test trial was finished, participants received feedback. The full trivia fact was presented with the answer bolded for a minimum of 4 seconds. After 4 seconds, a "Done" button would appear, and participants advanced to the next trial by clicking the button. In the 0-second control condition, the test trial was skipped and participants simply received a "feedback" trial. The answer was presented in bold so that participants in the 0-second condition would know which word would be left blank on the final test.

After the study phase, the distractor phase began. The distractor phase lasted two minutes. Participants were shown a text box with instructions to type as many countries as possible.

After the distractor phase, the final test phase began. Participants were asked the same questions as they had been asked (or had studied) in the first phase. The final test trials were self-paced and participants could advance to the next trial by pressing a button below the text box wherein they typed their responses.

After the final test phase, participants answered questions as to whether the experiment proceeded smoothly and if they had ever learned these materials before (participants indicating major problems with the experiment and/or prior participation in an experiment that used these facts were excluded from analyses). After the final questions were answered, the experiment ended. Participants were thanked for their participation and paid via Amazon's Mechanical Turk.

Table 1. Inferential statistics for overall effect of time spent in retrieval mode on initial test performance.

	df	*F*	*p*	η_p^2
Experiment 1a	3, 138	58.16	<.001	0.56
Experiment 1b	3, 144	50.05	<.001	1.04
Experiment 2a	2, 192	108.95	<.001	0.53
Experiment 2b	2, 134	81.93	<.001	0.55
Experiment 3	3, 168	54.31	<.001	0.49
Experiment 4	3, 126	43.30	<.001	0.51

Note: For Experiments 1a, 1b, 3, and 4, these analyses reflect comparisons between 0, 5, 10, or 30 seconds of retrieval. For Experiments 2a and 2b, these analyses reflect comparisons between 0, 5, or 10 seconds of retrieval.

Results

Initial test

The duration of the initial test (0, 5, 10, or 30 seconds) had a significant effect on the number of answers participants retrieved on that test (see Table 1). The more time participants were given, the more likely they were to retrieve the correct answer (see Figure 1). The same pattern was obtained when only including the retrieval conditions (5, 10, and 30 seconds), providing evidence that additional retrieval time was fruitful in terms of initial performance (see Table 2).

Final test

Our primary question was whether more time spent retrieving on the initial test would lead to higher recall on the final test (recall results are depicted in Figure 1). Overall, initial test duration (0, 5, 10, or 30 seconds) significantly affected final test performance (see Table 3). However, it was not the duration of the retrieval attempt that affected learning per se: recall on the final test did not differ significantly

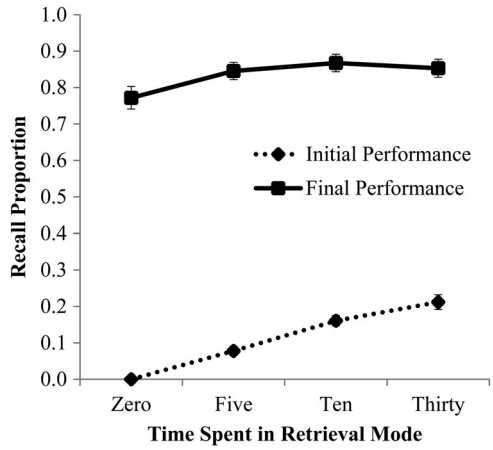

Figure 1. Proportion of items recalled on the initial and final tests in Experiment 1a as a function of the amount of time spent in retrieval mode (in seconds). Error bars report standard error of the mean.

among the 5-, 10-, and 30-second conditions (see Table 4). What did affect learning was whether or not a retrieval attempt was made at all: recall was significantly higher on the final test for items assigned to the 5-, 10-, and 30-second retrieval conditions ($M = .84$, $SD = .15$) than items assigned to the 0-second read-only condition ($M = .76$, $SD = .21$) (see Table 5), replicating the standard testing effect.

In short, giving participants more time to retrieve during the learning phase led to more retrieval success during the learning phase but not on the final test. Final test performance was affected by whether retrieval happened at all. In terms of the two-stage framework, the results of Experiment 1a suggest that engaging in Stage 1 is crucial, but the duration of Stage 1 did not affect learning.

Experiment 1b

The primary purpose of Experiment 1b was to replicate and extend the outcomes of Experiment 1a using a two-day retention interval. Often, testing effects increase in magnitude as the delay between initial learning and the final test increases (although not when participants are given feedback, which they were in the present study; see Kornell, Bjork, & Garcia, 2011). By assessing final test performance after a two-day delay, our aim was to do what we could to detect a significant effect of time spent in retrieval mode on final memory performance (if such a relationship existed).

Table 2. Inferential statistics for more versus less time spent in retrieval mode on initial test performance.

	df	*F* or *t*	*p*	η_p^2 or *d*
Experiment 1a	2, 92	26.70	<.001	0.15
Experiment 1b	2, 96	24.32	<.001	0.34
Experiment 2a	96	6.93	<.001	0.75
Experiment 2b	67	7.33	<.001	0.96
Experiment 3	2, 112	18.65	<.001	0.25
Experiment 4	2, 84	21.94	<.001	0.34

Note: For Experiments 1a, 1b, 3, and 4, these analyses reflect comparisons between 5, 10, or 30 seconds of retrieval. For Experiments 2a and 2b, these analyses reflect comparisons between 5 or 10 seconds of retrieval.

Table 3. Inferential statistics for overall effect of time spent in retrieval mode on final test performance.

	df	*F*	*p*	η_p^2
Experiment 1a	3, 138	8.19	<.001	0.15
Experiment 1b	3, 144	4.04	009	0.08
Experiment 2a	2, 192	100.96	<.001	0.51
Experiment 2b	2, 134	23.11	<.001	0.26
Experiment 3—Knew	3, 168	6.11	001	0.10
Experiment 3—Naive	3, 168	5.83	001	0.09
Experiment 4	3, 126	22.01	<.001	0.34

Note: For Experiments 1a, 1b, 3, and 4, these analyses reflect comparisons between 0, 5, 10, or 30 seconds of retrieval. For Experiments 2a and 2b, these analyses reflect comparisons between 0, 5, or 10 seconds of retrieval.

Method

Participants

Forty-nine participants (26 female, 23 male; median age = 33 years, range = 20–56 years) completed the experiment on Amazon's Mechanical Turk. Participants were compensated $4.00 for completing the first session and an additional $1.50 for completing the final test. All participants reported being fluent English speakers living in the US.

Materials and procedure

The materials and procedure used were identical to those used in Experiment 1a, except the final test occurred after a 48-hour delay.

Results

We conducted the same analyses in Experiment 1b as in Experiment 1a, and the results were the same in all important respects (see Figure 2). When participants were given more time on the initial test, they retrieved more answers to those questions (see Tables 1 and 2). In addition, there was a significant effect of initial test duration (0, 5, 10, or 30 seconds) on final test performance (see Table 3). However, the effect of initial test duration disappeared when we only considered the retrieval conditions (5, 10, and 30 seconds) (see Table 4). Thus, despite increasing initial test performance, more retrieval time did not ultimately enhance learning. What did affect learning was whether a retrieval attempt was made. Final test performance was significantly higher for items assigned to the retrieval conditions ($M = .64$, $SD = .20$) than for items assigned to the 0-second read-only condition ($M = .56$, $SD = .21$) (see Table 4), replicating the standard testing effect.

Table 4. Inferential statistics for more versus less time spent in retrieval mode on final test performance.

	df	*F* or t	*p*	η_p^2 or *d*
Experiment 1a	2, 92	0.75	.48	0.02
Experiment 1b	2, 96	0.50	.61	0.01
Experiment 2a	96	1.72	.09	0.18
Experiment 2b	67	1.12	.27	0.14
Experiment 3—Knew	2, 112	2.85	.06	0.05
Experiment 3—Naive	2, 112	2.66	.07	0.05
Experiment 4	2, 84	1.01	.37	0.02

Note: For Experiments 1a, 1b, 3, and 4, these analyses reflect comparisons between 5, 10, or 30 seconds of retrieval. For Experiments 2a and 2b, these analyses reflect comparisons between 5 or 10 seconds of retrieval.

Table 5. Inferential statistics for overall effect of retrieval versus presentation on final test recall.

	df	*t*	*p*	*d*
Experiment 1a	46	3.92	<.001	0.62
Experiment 1b	49	3.54	.001	0.51
Experiment 2a	96	12.87	<.001	1.43
Experiment 2b	67	6.19	<.001	0.76
Experiment 3—Knew	56	3.13	.003	0.46
Experiment 3—Naive	56	3.11	.003	0.46
Experiment 4	42	7.00	<.001	1.08

Note: For Experiments 1a, 1b, 3, and 4, these analyses reflect comparisons between retrieval (collapsed across 5, 10, or 30 seconds) to the read-only control condition. For Experiments 2a and 2b, these analyses reflect comparisons between retrieval (collapsed across 5 or 10 seconds) to the read-only control condition.

Feedback time

Across Experiments 1a and 1b, we analysed how long participants spent with feedback visible before they advanced to the next trial by clicking the button labelled "Done". Because some trials were outliers (i.e., very long reaction times), we first computed a median reaction time for each participant, and then averaged across those median values. Participants spent a relatively similar amount of time processing the feedback in the 0-second read-only condition ($M = 6.9$ seconds, $SD = 2.6$), the 5-second test condition ($M = 6.6$ seconds, $SD = 2.2$), the 10-second test condition ($M = 7.2$ seconds, $SD = 2.8$), and the 30-second test condition ($M = 7.1$ seconds, $SD = 2.0$). A repeated-measures ANOVA revealed a significant main effect of time in retrieval mode on total read times, $F(3, 285) = 3.86$, $p = .010$, $\eta_p^2 = 0.039$. On average, people spent more time processing the feedback in the longer retrieval mode conditions, which, if anything, should have helped them learn the information better but did not.

Discussion of experiments 1a and 1b

Attempting to retrieve enhanced learning when compared to the read-only (i.e., 0-second) control condition. However, more time in retrieval mode did not lead to better final test performance. This finding is surprising: more effort did not translate into better final test performance—and the evidence suggests that more effort was occurring, otherwise performance would not have improved during the initial study phase, which seems inconsistent with the retrieval effort hypothesis (Pyc & Rawson, 2009).

With respect to the two-stage framework (Kornell et al., 2015), the results from Experiments 1a and 1b suggest that: (a) completing both Stage 1 (i.e., the retrieval attempt) and Stage 2 (i.e., the processing of feedback) enhanced learning more than just completing Stage 2, and (b) the amount of time spent in Stage 1 does not seem to matter. These findings led us to a related question: does the amount of time spent in Stage 2 matter? Experiments 2a and 2b were designed to extend the results from Experiments 1a and 1b by testing this question. We manipulated two variables: time spent in Stage 1 and time spent in Stage 2.

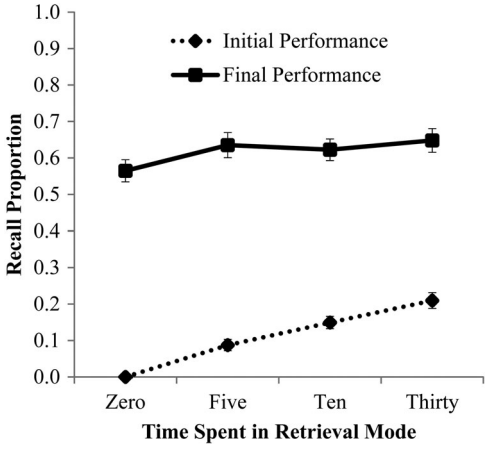

Figure 2. Proportion of items recalled on the initial and final tests in Experiment 1b as a function of the amount of time spent in retrieval mode (in seconds). Error bars report standard error of the mean.

Experiment 2a and 2b

In Experiments 2a and 2b, participants were given either 0, 5, or 10 seconds to retrieve an answer to a trivia question and either 2 or 7 seconds of feedback time to process the correct answer. The final test occurred either immediately (Experiment 2a) or after 48 hours (Experiment 2b).

We chose 2 or 7 seconds of feedback time based on the same principle that led us to choose 0, 5, 10, or 30 seconds in in Experiment 1: we tried to include durations most likely to capture an effect of duration if there is one. We felt that less than 2 seconds would be too short for meaningful processing. We also felt that if more time was going to be helpful, an additional 5 seconds should capture most of that value, and that there would probably be diminishing returns that would be small after 7 seconds. With respect to the timing of initial study, we removed the 30-second condition in Experiment 2 because a 2×4 design with 60 items would have meant there were relatively few items (7.5) per condition, whereas a 2×3 design allowed for 10 items per condition.

For items that were not tested, in the 0-second test condition, we predicted that final recall would be higher when items were presented for 7 seconds rather than 2 seconds. The prediction was not as obvious for items that were tested, in the 5- or 10-second condition, however. Perhaps more time in Stage 2 would help if Stage 1 was terminated before the search for the answer was complete. It may be possible to compensate for a shortage of retrieval time in Stage 1 with additional processing time in Stage 2. In this case, 7 seconds should be better than 2 seconds and we might predict that the relative benefit of 7 versus 2 seconds of feedback would be greater following 5 seconds of retrieval than 10 seconds of retrieval.

Another possibility is that as long as Stage 1 and Stage 2 occur (i.e., a retrieval attempt is made and feedback is given), learning will be equivalent regardless of the

amount of time participants spend processing the feedback. Experiments 1a and 1b demonstrated that the length of the retrieval attempt does not matter, despite the fact that active retrieval processing was occurring. It is possible that the essential processing necessary to learn from feedback happens relatively quickly in Stage 2, and as long as enough time is given to read the feedback, additional time processing the answer will not be beneficial.

Experiment 2a

Method

Participants

Ninety-seven participants (52 female, 45 male; median age = 33 years, range = 19–69 years) completed the experiment on Amazon's Mechanical Turk. Participants were compensated $4.00 for completing the experiment. All participants reported being fluent English speakers living in the US (except for one participant who did not answer the fluency question and another participant who did not clearly indicate his country of residence; both were included in the analyses).

Materials, design, and procedure

The materials used were identical to those used in Experiment 1. The procedure differed from Experiment 1 in two key ways. First, the 30-second condition was eliminated, leaving three test conditions (the 0-, 5-, and 10-second test conditions). Second, in Experiment 1 the answer (i.e., feedback) was presented for a minimum of 4 seconds, and then the participant pressed a button to move to the next trial. Feedback timing changed in Experiment 2a and 2b: Participants were given 2 seconds to process the feedback for half of the items and 7 seconds for the other half. Therefore, Experiment 2a and 2b used a 3 (test duration: 0, 5, or 10 seconds) × 2 (feedback duration: 2 or 7 seconds) design. As in the prior experiments, all manipulations occurred within-participants.

Results

Retrieval time

The design of Experiment 2a allowed us to repeat the analyses from Experiment 1a and 1b and examine the effect of retrieval time on final test performance. In our first set of analyses, we collapsed our data across the 2- and 7-second feedback conditions. As in Experiments 1a and 1b, when participants were given more time on the initial test, they retrieved significantly more answers on the initial test (see Figure 3, Tables 1 and 2). However, higher initial retrieval success was not associated with better final test performance (see Figure 3), thus replicating the previous experiments. Overall, initial test duration (0, 5, or 10 seconds) significantly affected final test performance (see Table 3), but the effect disappeared when considering only the two retrieval conditions (and not the read-only control condition). Ten seconds of initial retrieval time did not lead to better final test performance as compared to 5 seconds of initial retrieval time (see Table 4). Whether or not a retrieval attempt was made did matter. Items that were assigned to be retrieved (with the 5- and 10-second conditions combined) were recalled at higher rates on the final test ($M = .81$, $SD = .16$) than items assigned to 0-second read-only control ($M = .64$, $SD = .22$) (see Table 5).

Feedback time

The novel question in Experiment 2a was whether more time in Stage 2 (i.e., longer feedback) enhanced learning. A 3 (test duration: 0, 5, or 10 seconds) × 2 (feedback duration: 2 or 7 seconds) repeated-measures ANOVA revealed a main effect of test duration, $F(2, 192) = 86.48$, $p < .001$, $\eta_p^2 = 0.47$, which is consistent with the previous analyses on the effect of retrieval time. The key finding was that there was also a main effect of feedback duration, $F(1, 96) = 29.79$, $p < .001$, $\eta_p^2 = 0.24$. Performance on the final test was always better when 7 seconds of feedback followed the initial test than 2 seconds of feedback, though the effect was numerically small (see Figure 3). Finally, there was a significant interaction between test duration and feedback duration, $F(2, 192) = 3.29$, $p = .04$, $\eta_p^2 = 0.03$, providing evidence that the benefit of longer feedback depended on the duration of the retrieval attempt. In particular, the benefit of 7 seconds of feedback over 2 seconds of feedback was numerically largest in the 0-second read-only control condition. Follow-up paired samples *t*-tests revealed that the benefit of 7 seconds of feedback compared to 2 seconds of feedback was significant in the 0-second read-only control condition and the 10-second test condition (both $ps < .012$). This difference did not reach statistical significance in the 5-second test condition ($p = .074$).

In sum, Experiment 2a replicated the findings of Experiment 1 but went further: spending more time in Stage 1 did not enhance learning, but spending more time in Stage 2 did. Experiment 2b was designed to

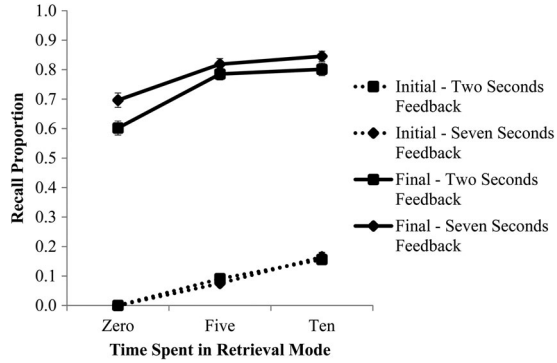

Figure 3. Proportion of items recalled on the initial and final tests in Experiment 2a as a function of the amount of time spent in retrieval mode (in seconds) and the amount of feedback time given. Error bars report standard error of the mean.

replicate and extend the results from Experiment 2a across a 2-day delay.

Experiment 2b

Method

Participants

Sixty-eight participants (48 female, 20 male; median age = 33 years, range = 19–63 years) completed the experiment on Amazon's Mechanical Turk. Participants were compensated $4.00 for completing the first session and an additional $1.50 for completing the final test. All participants reported being fluent English speakers living in the US (except for one participant who did not answer the fluency question; this person was included in the analyses).

Materials and procedure

The materials and procedure used were identical to those used in Experiment 2a, except the final test occurred after a delay of 48 hours.

Results

Retrieval time

Consistent with all of the previous experiments, Experiment 2b revealed that additional time on the initial test led to more retrieval success on the initial test (see Figure 4, Tables 1 and 2). However, the benefits of additional retrieval time did not transfer to the final test. Although there was a significant effect of initial test duration overall (0, 5, and 10 seconds) on final test performance (see Table 3), the effect disappeared when we considered only the retrieval conditions. An initial retrieval attempt of 10 seconds did not lead to higher recall on the final test as compared to 5 seconds (see Table 4). Thus, it was not retrieval time that affected recall per se, it was whether or not a retrieval attempt was made. Items that were assigned to be retrieved initially (for 5 or 10 seconds) were recalled significantly more often on the final test ($M = .68$, $SD = .18$) than items assigned to the 0-second read-only control condition ($M = .56$, $SD = .21$; see Table 5).

Feedback time

The primary question in Experiment 2b was whether more time in Stage 2 (i.e., longer feedback) enhanced learning, with the final test occurring after 48 hours. The results of Experiment 2b mostly replicated those of Experiment 2a. A 3 (test duration: 0, 5, or 10 seconds) × 2 (feedback duration: 2 or 7 seconds) repeated-measures ANOVA revealed a main effect of test duration, $F(2, 132) = 15.95$, $p < .001$, $\eta_p^2 = 0.20$. Critically, there was also a main effect of feedback duration, $F(1, 66) = 23.41$, $p < .001$, $\eta_p^2 = 0.26$. Performance on the final test was always better when 7 seconds of feedback followed the initial test compared to 2 seconds of feedback (see Figure 4). Finally, unlike Experiment 2a, there was no significant interaction between test duration and feedback duration, $F(2, 132) = 2.21$, $p = .11$, $\eta_p^2 = 0.03$, suggesting that the benefit of longer feedback did not depend on the duration of the initial retrieval attempt.

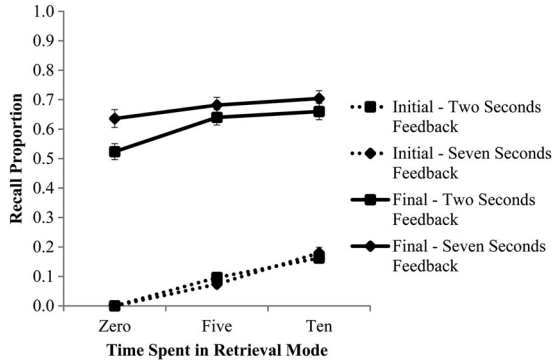

Figure 4. Proportion of items recalled on the initial and final tests in Experiment 2b as a function of the amount of time spent in retrieval mode (in seconds) and the amount of feedback time given. Error bars report standard error of the mean.

Discussion of experiments 2a and 2b

Experiments 2a and 2b served two functions. First, they allowed us to replicate the findings of Experiment 1a and 1b. Consistent with these prior experiments, we found more retrieval success following longer versus shorter retrieval attempts. However, increased success on the initial test did not translate into increased recall on the final test, whether the final test was immediate or after a 48-hour delay.

Experiments 2a and 2b also extended the results of Experiments 1a and 1b by testing the effect of feedback duration. In both experiments, 7 seconds of feedback led to significantly higher recall on the final test than 2 seconds of feedback, though the difference was numerically small. In Experiment 2b, the benefit of longer feedback did not depend on the length of the retrieval attempt, but it did in Experiment 2a. This interaction was driven by a larger difference between the effect of 2 and 7 seconds of feedback in the 0-second read-only control condition than in the 5- or 10-second retrieval conditions (see Figure 3). For items in the 0-second read-only control condition receiving 2 seconds of feedback, recall was particularly low on the final test because total time to study an item in this condition was only 2 seconds (and this may not have been enough time to fully encode the question and answer).

Taken together, the first two experiments support two main conclusions: learning benefits from spending more time processing feedback (in Stage 2), but not from spending more time in retrieval mode (in Stage 1).

Experiment 3

Although Experiments 1 and 2 were about time spent trying to retrieve, they led us to a different question:

does retrieval success enhance learning? This question came up because differences in initial test performance were not associated with differences in final test performance; in other words, more retrieval success did not seem to produce more learning. Experiment 3 examined the effect of retrieval success on learning (to foreshadow, Experiment 3 suggests that none of the studies we report here should be seen as evidence either way about whether retrieval success matters).

Whether or not retrieval success matters, in the paradigm we are using, boils down to a question about a key subset of items: items that the participant knew going into the experiment. For these items, unlike items participants did not know, more time trying to retrieve can lead to more retrieval success during the initial study phase. We wanted to focus our analysis on these items in order to examine the effect on learning given differences in retrieval success. However, we cannot simply compare items participants did not recall on the initial test to ones they did because of item selection effects. That is, items that are recalled on the initial test are easier items for that participant than those not recalled on the initial test; therefore, directly comparing recalled to non-recalled items conflates retrieval success with item difficulty.

In Experiment 3, we used a methodology that allowed us to identify the subset of items participants already knew. Participants were asked to press a button labelled, "Knew this already—continue" if they thought they knew the answer or to press a button labelled, "Did not already know this—continue" if they did not think they knew it. These buttons were available to participants on the feedback trial, after the item had been initially tested. We could then focus our analysis on items that participants said they knew, or had already answered correctly, and exclude items they did not know. As such, we can examine the extent to which retrieval success in the longer retrieval conditions (e.g., 30 seconds) influenced performance relative to the absence of retrieval success in the shorter retrieval conditions (e.g., 5 seconds) for items that participants knew (i.e., for items that were roughly equivalent in difficulty). Additionally, by assessing which items participants did not already know, we can examine the extent to which more versus less time in retrieval mode influenced final performance for novel items.

This method is not valid if self-reports are affected by the amount of time participants spent in retrieval mode (or the associated differences in retrieval success). It assumes that items participants said they knew in any one condition are equivalent to items they said they knew in other conditions. We test this assumption, in the results section, by examining the overall rate at which participants said they knew items across conditions (and find that it is valid).

Methods

Participants

Fifty-seven participants (36 female, 21 male; median age = 32 years, range = 19–56 years) completed the experiment on Amazon's Mechanical Turk. Participants were compensated $4.00 for completing the experiment. All participants reported being fluent English speakers living in the US. A fifty-eighth participant, who said she was not a fluent English speaker and did not answer the question about living in the US, was excluded from the following analyses.

Materials and procedure

The materials were the same as the ones used in Experiment 1 and 2. The procedure was nearly identical to the one used in Experiment 1a: participants were given 0, 5, 10, or 30 seconds to retrieve, feedback duration was under participants' control after a minimum of 4 seconds, and the final test was immediate. The only difference was how participants ended feedback trials to advance to the next trial. In every experiment, feedback was the question and answer being shown together. In Experiment 1a, participants clicked a "Done" button. In Experiment 3, there were two buttons: "Knew this already—continue" and "Did not already know this—continue". As in Experiment 1a, feedback time lasted a minimum of 4 seconds, after which the two buttons appeared and advanced to the next trial upon being clicked. In short, the timing of Experiment 1a and Experiment 3 were the same and the only difference between them was that in Experiment 3 participants indicated whether or not they already knew the item during feedback.

Results

Retrieval time

As in all previous experiments, when participants were given more time to retrieve on the initial test, they produced more correct answers (see Figure 5, Tables 1 and 2). Initial test duration (0, 5, 10, and 30 seconds) also significantly affected recall on the final test (see Table 3). However, final test performance was similar among the retrieval conditions (5, 10, and 30 seconds) (see Table 4), indicating that more retrieval time did not enhance learning. The p value for this test was .06, suggesting that perhaps there was at least a small effect of retrieval time on final performance. However, the effect size was small (.05) and final test performance did not increase monotonically with initial test duration ($M = .83$, $SD = .18$; $M = .87$, $SD = .16$; $M = .85$ $SD = .15$ in the 5-, 10-, and 30-second retrieval conditions, respectively). Thus, this variation in final test performance is not strong evidence for the idea that retrieval success affects learning.

Whether a retrieval attempt was made did reliably affect learning, though. Final test performance was significantly higher for items in the retrieval conditions ($M = .85$, $SD = .15$) than the 0-second read-only control condition ($M = .80$, $SD = .20$) (see Table 5).

Feedback time

As in Experiments 1a and 1b, we analysed how long participants processed the feedback after the test trial (as a

reminder, they had to advance to the next trial manually by clicking either the "Knew this already—continue" button or the "Did not already know this—continue" button). Again, we computed a median reaction time for each participant, and then averaged across those median values. Participants spent a relatively similar amount of time processing the feedback in the 0-second read-only condition ($M = 6.4$ seconds, $SD = .8$), the 5-second test condition ($M = 5.9$ seconds, $SD = 1.1$), the 10-second test condition ($M = 6.3$ seconds, $SD = 1.5$), and the 30-second test condition ($M = 6.5$ seconds, $SD = 1.7$). A repeated-measures ANOVA revealed a significant main effect of time in retrieval mode on total read times, $F(3, 126) = 2.99$, $p = .034$, $\eta_p^2 = 0.066$. However, as in Experiments 1a and 1b, people tended to spend more time processing the feedback in the longer retrieval mode conditions (which, once again, should have helped them learn the information even better but did not).

Retrieval success

Participants said that they already knew the answer at similar rates after 0, 5, 10, and 30 seconds of retrieval (on 25%, 24%, 25%, and 23% of trials, respectively). These rates justify the assumption that self-reports were not affected by the amount of time allowed for retrieval. In the next analysis, we only consider the items participants said they knew, because these are the items for which more time could lead to more retrieval success. Among these items, the four retrieval time conditions differ in terms of retrieval success, but not item difficulty. Therefore, we can compare final test performance to determine whether retrieval success enhances learning with minimal concern for item selection effects.

If retrieval success enhances learning, we would expect to see longer initial test durations lead to higher recall on the final test. However, our results do not allow us to detect an effect of retrieval success because of a ceiling effect. Among the items participants indicated that they already knew, final test performance was above 95% in all four retrieval time conditions (see Figure 5), which did not differ significantly from one another, $F(3, 147) = 1.01$, $p = .39$, $\eta_p^2 = 0.02$. Although an effect of retrieval success may exist and may emerge in some situations, we were not able to detect the effect in this experiment due to near perfect recall performance across all conditions. We elaborate on the implications of this finding in the discussion.

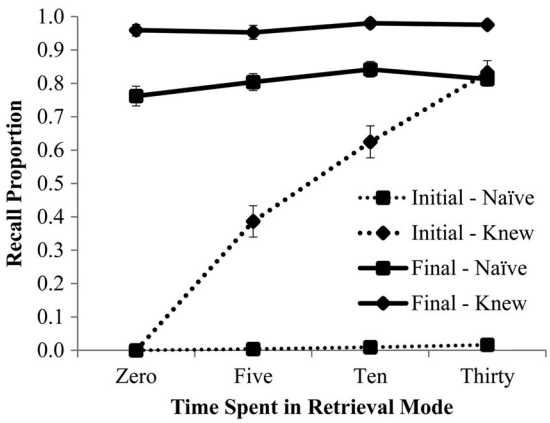

Figure 5. Proportion of items recalled on the initial and final tests in Experiment 3 as a function of the amount of time spent in retrieval mode (in seconds) and prior knowledge. Error bars report standard error of the mean.

New learning

Finally, we assessed the effect of additional retrieval time on new learning by examining items that participants indicated they did not already know. Participants rarely guessed correctly on the initial test on these items (approximately 1% of the time). Yet on the final test, participants recalled an average of 76%, 80%, 84%, and 81% of these "did not know" items in the 0-, 5-, 10-, and 30-second conditions, respectively.

Among these items, the same pattern of results was observed as in prior experiments. There was an overall main effect of initial test duration (see Table 3). However, more time spent making a retrieval attempt on the initial test did not appear to make an answer more memorable on the final test as the 5-, 10-, and 30-second conditions did not differ significantly on the final test (see Table 4). Even though participants guessed incorrectly on the "did not already know" items when making a retrieval attempt, a testing effect emerged. Items that were assigned to a retrieval attempt of any duration were recalled more frequently ($M = .82$, $SD = .16$) than items assigned to the 0-second read-only control condition ($M = .76$, $SD = .22$) (see Table 5).

Discussion of experiment 3

In Experiment 3, we asked whether retrieval success influenced learning. The results did not provide an answer to this question. The main finding was that our paradigm precluded the possibility of retrieval success having an effect, because performance on items that participants already knew was near the ceiling on the final recall test. This finding is important because without it, the reader might come away with the false impression that our data indicate retrieval success does not matter. Our data do not say anything about the effect of retrieval success.

However, the data have a lot to say about our main question, which is the effect of retrieval time on learning. Experiment 3 replicated the previous studies with a twist, this time examining items that participants did not already know. Like in the previous studies, a testing effect emerged (in this case, it was better to attempt retrieval and fail than to not attempt retrieval and only study the information). However, no differences emerged in final test performance for more versus less time spent in retrieval mode for these novel items.

The finding that testing improved memory for items that participants did not already know (and therefore

failed to retrieve on the initial test approximately 99% of the time) is consistent with prior research showing pretesting effects (e.g., Kornell, Hays, & Bjork, 2009). In these paradigms, participants were pretested (e.g., whale-????) before being presented with an initial study phase (e.g., whale-mammal), rendering the items non-retrievable. Yet, pretesting often enhanced memory more than a read-only control condition, despite the fact that the rate of success, which occurred because of guessing, was very low (and any items that were successful on the pretest were excluded from analyses). Other researchers have also found benefits of pretesting (e.g., Grimaldi & Karpicke, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, 2014; Vaughn & Rawson, 2012). As many authors have pointed out, the difficulty in examining the effect of unsuccessful retrieval is identifying items that participants do not know. Experiment 3, which identified such items based on self-report, adds to this body of research. It is especially valuable because it used questions about true information that allowed participants to do a meaningful memory search (some studies have basically involved guessing rather than normal retrieval). Furthermore, in most studies items that participants answer correctly have to be excluded from the analyses, which biases the results against the retrieval condition, but the method in Experiment 3 did not require us to exclude any items from the analysis, providing a more accurate picture of the magnitude of the effect of unsuccessful retrieval.

Turning to the effect, or lack thereof, of retrieval time on learning in Experiment 3, one potential concern is that participants may have terminated their search quickly for novel items. For instance, participants presented with the question, "Dominos were invented in the country of _____", may realise rather quickly that they do not know the correct answer. If this realisation occurs quickly, participants may simply end their search for the correct answer, which would be problematic as time spent in retrieval mode was the primary manipulation in these experiments. In other words, one could argue that there are relatively uninteresting reasons why time in retrieval mode did not matter: for items that participants already knew, there was a ceiling effect, and for items they did not already know they did not spend more than 5 seconds searching their memory. We are not saying this argument is correct, but we decided to conduct Experiment 4 to address this issue. Note that even if this were the case, it does not change the practical implications of this finding: among a mixture of known and unknown items, being given more time to try to retrieve the answers does not enhance learning.

Experiment 4

Experiment 4 used learning materials taken from the remote associates task (RAT), in which participants are shown three words and have to think of a fourth word that is related to the other three. It was unlikely that participants would have been exposed to any of the problems before starting the study. Moreover, with trivia questions it is often possible to make a quick decision that trying to answer is futile, but this is not true with RAT problems; there is always a chance that continuing to try will pay off. Furthermore, these problems are tantalising and tend to elicit a strong motivation to find a solution.

Methods

Participants

Forty-three participants (25 female, 16 male, 2 unreported; median age = 34 years, range = 22–62 years, 1 unreported) completed the experiment on Amazon's Mechanical Turk. Participants were compensated $4.00 for completing the experiment. All participants reported being fluent English speakers living in the US (except for one participant who did not answer the question about English fluency; this person was included in the analyses).

Materials

The materials were 60 remote associates test (RAT) problems taken from the norms of Bowden and Jung-Beeman (2003). These problems involve three seemingly unrelated words (e.g., *teeth*, *arrest*, and *start*) that are related through a fourth word (e.g., *false*). Other example RAT problems include: *artist*, *hatch*, *route* (solution: *escape*); and *reading*, *service*, *stick* (solution: *lip*). A full list of the RAT stimuli used in Experiment 4 is reported in Appendix B.

Procedure

The procedure was identical to the one used in Experiment 1a, with the only exception being the materials used. On test trials, participants were shown three remote associates (e.g., *self*, *attorney*, and *spending*) and were given a text box to type the target word that related the three remote associates (e.g., *defense*). On feedback trials and on the 0-second read-only trials, the three remote associates were presented along with the correct answer (which was presented in a bold font).

Results

Initial test

When participants were given more time on the initial test, they solved significantly more RAT problems (see Figure 6). This was true whether we included all four time conditions (0, 5, 10, and 30 seconds) (see Table 1) or just the retrieval conditions (5, 10, and 30 seconds) (see Table 2).

Final test

The key question was whether recall on the final test would also benefit from more time on the initial test. As with trivia questions, it did not. Although there was an overall effect of initial test duration (0, 5, 10, and 30 seconds) on final test performance (see Table 3), the effect did not remain when we considered only the retrieval conditions (see

Table 4). However, we did replicate the testing effect. Attempting to solve a RAT problem (for 5, 10, or 30 seconds) before being shown the answer ($M = .65$, $SD = .19$) benefitted memory more than being shown the answer immediately as in the 0-second read-only condition ($M = .46$, $SD = .21$) (see Table 5).

Feedback time

As in Experiments 1a and 1b, we analysed how long participants processed the feedback after the test trial (as a reminder, they had to advance to the next trial manually by clicking a button labelled "Done"). In order to control for outliers, we first computed median reaction times for each participant, and then averaged across those median values. Participants spent a relatively similar amount of time processing the feedback in the 0-second read-only condition ($M =$ 6.0 seconds, $SD = 1.8$), the 5-second test condition ($M = 6.4$ seconds, $SD = 2.1$), the 10-second test condition ($M = 6.9$ seconds, $SD = 3.1$), and the 30-second test condition ($M =$ 7.0 seconds, $SD = 2.7$). A repeated-measures ANOVA revealed a significant main effect of time in retrieval mode on total read times, $F(3, 126) = 2.99$, $p = .034$, $\eta_p^2 = 0.066$, and on average people tended to spend more time processing the feedback in the longer retrieval mode conditions (which, once again, should have helped them learn the information even better but did not).

Discussion of experiment 4

The results of Experiment 4 replicated the findings of the previous studies using a RAT problem-solving task instead of trivia questions. Experiment 4 is also useful because it provides reassurance about a potential criticism of Experiments 1–3. We speculated in the discussion of Experiment 3 that for a subset of items, our participants quickly decided they did not know the answer (Son & Metcalfe, 2005) and stopped trying within 5 seconds, making our retrieval time manipulation ineffective. This criticism might leave open the possibility that more retrieval time really does enhance learning. If this was true, though, then in Experiment 4—where it seems unlikely that participants gave up, because thinking of the answer is always both possible and alluring—there should have been an effect of providing participants with more time to think of the answer. The fact that there was no such effect strengthens the claim that as long as one attempts to retrieve, spending more or less time in retrieval mode affects retrieval success without affecting learning.

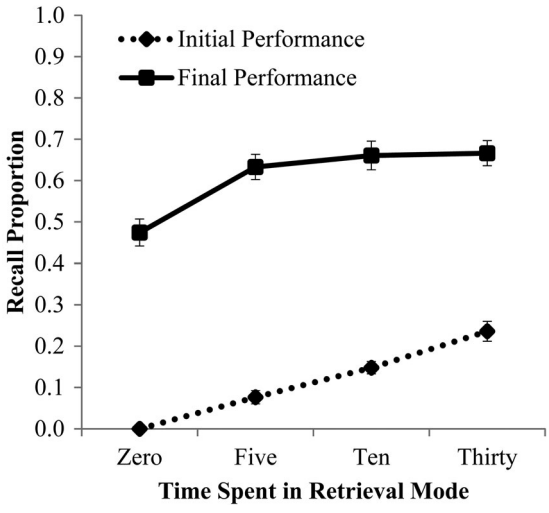

Figure 6. Proportion of items recalled on the initial and final tests in Experiment 4 as a function of the amount of time spent in retrieval mode (in seconds). Error bars report standard error of the mean.

General discussion

A consistent, albeit novel and unexpected, pattern emerged across four experiments. Retrieval attempts enhanced learning, and spending more time trying to retrieve increased initial retrieval success, but the amount of time spent in retrieval mode did not affect *learning*. This finding is intriguing for several reasons. At a general level, the large amount of prior evidence supporting the desirable difficulties framework (e.g., Bjork, 1994) would seem to predict that more active processing and effort would have enhanced learning. The findings also have implications for two process-based theories, the retrieval effort hypothesis and the elaborative retrieval hypothesis, which we discuss after re-addressing a potential concern with the present experiments.

Revisiting terminating search early

As mentioned before, one potential criticism of the current work is that participants may have terminated their search early (i.e., before the allotted time had elapsed). Early termination of the search would have unintentionally equated our conditions, such that participants in the 30-second condition would not have spent more time trying to retrieve than participants in the 5-second condition. The evidence weighs against this concern. If participants had terminated their search early, then the level of retrieval success should have been approximately equal in the 5-, 10-, and 30-second retrieval conditions, but it was not. Participants recalled more correct answers during the initial practice session when given more time to retrieve. Additionally, the same increase in initial performance was observed for both the RAT materials and trivia questions, providing further evidence that participants were not terminating their search early for most items. We do not claim to know the exact nature of the retrieval processing in which they were engaged, but it does appear that participants spent more time in retrieval mode when they were given more time to retrieve.

Feedback timing

Our manipulation of time in retrieval mode bears more than a passing resemblance to prior research on

immediate versus delayed feedback. In our experiments, we presented participants with a problem and told them the correct answer after a variable amount of time (e.g., 5, 10, or 30 seconds). If we presume, for a moment, that participants always stopped making a retrieval attempt after 5 seconds, then presenting the correct answer after 5 seconds could be considered immediate feedback, whereas presenting the correct answer after a longer delay (e.g., 10 or 30 seconds) could be considered delayed feedback. Prior research has shown memory benefits of delayed feedback (e.g., Butler, Karpicke & Roediger, 2007; Metcalfe, Kornell, & Finn, 2009; see also Kulik & Kulik, 1988).

However, it is almost certainly untrue that participants always stopped trying to retrieve after 5 seconds and, thus, the current methodology differs substantially from those typically used to investigate timing of feedback effects. Delayed feedback is usually manipulated by having participants take a test on an item (e.g., "horse-????"), take tests on an intervening set of different items (e.g., "ghost-????" and "forest-????"), and then receive feedback on the original item (e.g., "horse-walnut"). Increasing the delay does not give participants more time to try to retrieve the answer. In the current experiments, by contrast, participants were tested on the original item (e.g., "Dominos were invented in the country of _____"), and then (presumably) continued to try to retrieve the answer (e.g., China) to that item for either 5, 10, or 30 seconds. Thus, in contrast to studies examining feedback timing, in the current methodology increasing the delay between the question and answer gave participants more time to think of the answer. In summary, research on delayed feedback differs in a crucial way from the research presented here, and it would be inappropriate to make strong predictions about one based on the other. It is also worth noting, however, that delaying feedback enhances learning, so if there were important similarities between the paradigms, it would be all the more reason to expect participants to learn more if they spend more time in retrieval mode. Of course, this did not occur in the present experiments, despite the similarities to a delayed feedback paradigm.

Retrieval effort

The amount of cumulative retrieval effort differed across conditions, as participants spent more time making an effort to retrieve in some conditions than others. The retrieval effort hypothesis, at least in its current form (e.g., Pyc & Rawson, 2009), would predict that this increased effort should increase learning. The data showed otherwise. One way to modify the retrieval effort hypothesis would say the intensity of effort predicts the amount of learning, but the duration of the effort does not (at least after an initial effortful attempt is made).

A more sweeping revision of the theory might be in order. It is possible that retrieval effort per se does not actually matter. Retrieval effort effects come about when more difficult retrieval produces more learning. Perhaps the difficulty of the retrieval does not play a causal role in the amount of learning. Instead, perhaps retrieval effort effects come about because more difficult items have more room to be strengthened in memory (e.g., they have less retrieval strength or storage strength; Bjork & Bjork, 2011). In other words, perhaps retrieval effort effects should be explained based on item learnability, not retrieval effort itself. This item-learnability theory would explain why more difficult items benefit more from retrieval attempts (they have more room for learning to occur). It also fits easily with the current finding, that spending more time in retrieval mode did not affect learning, because the duration of the retrieval attempt did not affect the difficulty of the items.

Finally, we note that using more concrete measures of retrieval effort may benefit future research into this area. Research suggests that pupil dilation is indicative of overall mental effort (e.g., Koelewijn, Zekveld, Festen, Kramer, 2012). Measuring pupil dilation during learning could serve as a manipulation check for future experiments investigating the effects of retrieval effort on learning or to examine just how "difficult" a particular manipulation happens to be.

Elaborative retrieval

The current findings are intriguing for other reasons as well. According to the elaborative retrieval hypothesis (e.g., Carpenter, 2009, 2011), part of the reason retrieval is beneficial is that it increases the activation of related information in memory. For instance, Carpenter (2011) had participants either study or complete test practice with weakly related word pairs (e.g., Mother-Child). On a final test, Carpenter gave participants either the original cue word (e.g., Mother) or a semantic mediator as the cue word (e.g., Father) with instructions to recall the target word (e.g., Child). When participants were given the mediator cue on the final test (e.g., Father), Carpenter found that recall of the target word (e.g., Child) was much better in the practice testing condition versus the study condition. The idea is that only test trials activate related semantic information during learning (e.g., Mother is semantically related to Father, and thus gets activated during a test trial) that can later be used to recall target information.

In our experiments, things like "World War II", "Germany", and "leader" might be activated by the question, "Who was Time Magazine's 'Man of the Year' in 1938?" This additional activation facilitates the formation of mediated connections from the cue to related information to the target ("Hitler"). In the current studies, one might assume that a longer time spent in retrieval mode would correspond to an increased amount of semantic activation, which would enhance learning. This kind of active processing is recommended by educators and researchers alike, and more elaborative processing during

encoding is often associated with enhanced memory performance (e.g., Bjork, 1994; Bjork & Bjork, 2011; Bradshaw & Anderson, 1982; Fisher & Craik, 1980; Schacter & Graf, 1986). Our findings are inconsistent with a simple conception of elaborative processing as a path to learning, given that any additional active processing beyond 5 seconds was not associated with improved final memory performance. We speculate below as to why this might be the case.

One might try to resolve this apparent conflict by claiming that the information activated during a retrieval attempt quickly reaches an asymptote. During the first few seconds, a participant might activate the vast majority of their to-be-activated semantic information. As time progresses, diminishing returns would ensue and increasingly less semantic information would be activated. If true, then the process of activating related information, which is central to the elaborative retrieval hypothesis, would only occur during the first few seconds, and additional time spent in retrieval mode would not be assumed to be useful. However, one potential flaw with this reasoning stems from the fact that initial performance increased as time in retrieval mode increased, which at least suggests that participants were continuously engaged in the retrieval process. However, whether or not this extra time spent in retrieval mode specifically increased the amount of semantic activation is unclear, as we did not directly measure the amount of semantic activation. If participants had been terminating their search after the first few seconds, then this pattern would not have obtained. This finding supports the idea that spending more time in retrieval mode did activate at least some new related information.

If our participants did, as we suspect, activate increasing amounts of related information as the amount of time in retrieval mode increased, then what are the implications for the elaborative retrieval hypothesis? The elaborative retrieval hypothesis states that the information activated at the time of the initial retrieval attempt mediates recall of the answer on subsequent attempts. However, it is reasonable to assume that not all information activated during the initial retrieval attempt is created equal in terms of its quality as a mediator. One possibility is that after the first few seconds, most of the key related concepts have been activated, and that the key concepts are the ones that are likely to serve as mediators. For example, asking a question like "_____ was the first state in the US to allow women to vote", should activate concepts like voting, states, the women's suffrage movement, revolutionary legal changes, and many specific states (e.g., Indiana, Colorado, California) within a few seconds. Less relevant (but related) information might be activated further along in time, such as hanging chads and voter ID cards. The less relevant information that is activated later might not be particularly useful as a mediator to the correct answer (and therefore is of little benefit according to the elaborative retrieval hypothesis). Not only could this tangential information make for low-quality mediators, it might actually provide a form of interference. For instance, activating the concept "hanging chad" might make one think of Florida, which could then interfere with the correct answer (in this case, Wyoming). This explanation is speculation but we have not come up with a better way of making sense of our findings based on the elaborative retrieval hypothesis. Future research could further investigate the issue of quantity versus quality of mediators.

Time retrieving versus time processing feedback

In contrast to the total time spent in retrieval mode, the length of time participants spent studying the correct answer did seem to matter. In Experiment 2, participants remembered the solution more often when they had studied the correct answer for 7 versus 2 seconds, regardless of how long they had spent in retrieval mode. More time leading to more learning does not seem like a surprising finding, except in light of the fact that more time trying to retrieve did not lead to more learning. Previous studies have shown that studying more improves learning (albeit with diminishing returns; e.g., Nelson & Leonesio, 1988) and that interrupting processing of the answer after a retrieval attempt can impair learning (e.g., Finn & Roediger, 2013). Extra time to process the solution may be beneficial for several reasons. Next we describe one possible explanation, which we call the *integration hypothesis*.

It is possible that learning requires having sufficient time such that the question processing done in Stage 1 can be integrated with the answer information provided in Stage 2. When feedback time was reduced to 2 seconds, participants may not have been able to fully integrate their activated information with the solution. The integration would presumably involve both understanding the answer and making connections in their existing memory network; if one adopts the elaborative retrieval hypothesis, one would assume that time is spent solidifying the direct link between the cue and target and simultaneously linking the cue to the target via mediated connections.

This integration hypothesis predicts that retrieval attempts will enhance learning if two critical conditions are met: (1) some minimum amount of time is allowed for the retrieval attempt and (2) sufficient time is allowed for processing the answer once it becomes available. Thus, as in the two-stage framework, the integration hypothesis states that retrieval will improve learning as long as both Stage 1 (the retrieval attempt) and Stage 2 (the processing of feedback) occur. Furthermore, for items of equivalent difficulty, it also identifies three features of a retrieval attempt that are not crucial: (1) the amount of time spent attempting to retrieve beyond the minimum, (2) the total amount of effort put into the retrieval attempt, and (3) the total amount of information activated during the retrieval attempt.

Educational implications

The results from one set of laboratory studies should not lead to universal recommendations for educators, especially considering the variability that exists inside the classroom with respect to materials, class size, and so on. We warn readers to be cautious about over-interpreting the educational implications of our findings, which we describe next.

One concern that educators may have is that testing takes too much time to implement in the classroom. Our results suggest that relatively brief quizzes or questions will enhance student learning. After asking a question, perhaps a teacher needs to pause only long enough for students make a retrieval attempt. The results also suggest, though, that it is crucial to pause after telling students the answer, so they can process it. For example, asking a question, pausing 10 seconds, providing the answer and then immediately moving on might be less effective than asking the question, pausing 5 seconds, providing the answer, and then pausing another 5 seconds (a similar comparison was made, in Experiment 2, by comparing a 5-second question with 7 seconds of feedback to a 10-second question with 2 seconds of feedback).

Furthermore, research has shown that covert retrieval affords the same benefits as overt retrieval (e.g., Smith, Roediger, & Karpicke, 2013). Thus, formal testing may not be necessary in the classroom in order to reap the benefits of retrieval practice. It is important to note, though, that there may be advantages of formal tests. First, students might not to try to retrieve when tested informally, whereas a formal test basically guarantees that students will make a retrieval attempt. Second, formal tests might increase motivation to study prior to class. Again, these comments are speculation.

Semantic versus episodic memory

The current experiments used general knowledge questions, and used a pretesting paradigm to assess whether more time in Stage 1 or Stage 2 improved learning. A pretesting paradigm taps into knowledge that has already been acquired (i.e., semantic memory). In contrast, we could have used a different set of materials (e.g., foreign language word pairs) and provided an initial study phase before the retrieval phase (i.e., testing episodic memory for newly learned facts). The distinction between the two memory systems is interesting, and has been noted in similar research before. For instance, Benjamin et al. (1998) highlight that general knowledge questions tap into semantic memory, but that free recall tests are primarily episodic. As a reminder, they found that taking longer to retrieve from semantic memory on an initial test was associated with better performance on a final episodic free recall test. To revisit, we found that taking longer to retrieve from semantic memory did not improve final performance on the same cued recall test later. But was our final test still tapping into semantic memory, episodic memory, or a mixture of both? We argue that it was a mixture of both, in the sense that participants could either: (1) try to remember the specific answer they learned during practice, when they received feedback, and/or (2) try to remember the answer from their general knowledge bank. Although we do not currently have a theoretical reason to suspect that retrieval operates differently in semantic versus episodic memory, it might make for an interesting line of future research in terms of examining how time in retrieval mode influences learning in the two memory systems.

Conclusion

The current results improve our understanding of retrieval effects in two ways. First, more time spent in retrieval mode does not confer additional benefits in final memory performance. Second, the amount of time processing the feedback should be given careful consideration both within a classroom setting and for students attempting to learn information outside of the classroom. Practically speaking, if students fail to retrieve the answer to a question after a few seconds, they should terminate their search and spend more time processing the correct answer. Spending additional time in retrieval mode will likely not help them remember the solution any better than if they had only attempted retrieval for a few seconds, even if they are exerting continual effort during the retrieval attempt. In other words, spinning your wheels while trying to retrieve does about as much good, for learning, as spinning your wheels when your car is stuck in the mud. However, spending a few extra seconds processing the correct answer seems to be a worthwhile time investment.

Notes

1. The Monkees.
2. In general, we prefer to err on the side of including as many participants as possible because doing so prevents potentially subjective decision making from having a large influence on data analyses. Thus, we included participants if they indicated that they lived in the US or indicated that they were fluent in the English language (and only excluded participants if they failed to give the desired answer to both of these questions). Across studies this policy resulted in the inclusion of 5/361 (1.4%) participants and the exclusion of one participant.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by a grant awarded to the third author by the James S. McDonnell Foundation [220020371].

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940–945.

Auble, P. M., & Franks, J. J. (1978). The effects of effort toward comprehension on recall. *Memory & Cognition*, *6*(1), 20–25.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55–68.

Benassi, V., Overson, C. E., & Hakala, C. (Eds.). (2014). *Applying science of learning in education: Infusing psychological science into the curriculum.* Retrieved from http://teachpsych.org/ebooks/asle2014/index.php

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society, making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning* (pp. 56–64). New York: Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge: MIT Press.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 634–639.

Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 165–174.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491–1494.

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.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*(6), 945–955.

Finn, B., & Roediger III, H. L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1665–1681.

Fisher, R. P., & Craik, F. I. (1980). The effects of elaboration on recognition memory. *Memory & Cognition*, *8*(5), 400–404.

Gardiner, F. M., Craik, F. I., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, *1*(3), 213–216.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513.

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

Kerr, R., & Booth, B. (1978). Specific and varied practice of motor skill. *Perceptual and Motor Skills*, *46*(2), 395–401.

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731–746.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*(2), 291–300.

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 106–114.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294.

Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79–97.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*(8), 1077–1087.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 676–686.

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.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498.

Schacter, D. L., & Graf, P. (1986). Effects of elaborative processing on implicit and explicit memory for new associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 432–444.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition*, *6*(4), 342–353.

Smith, M. A., Roediger III, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1712–1725.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*, *33*(6), 1116–1129.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon.

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*(5), 899–905.

Appendix A. Trivia fill-in-the-blank questions used in Experiments 1a, 1b, 2a, 2b, and 3

Question	Solution
The US state of _____ has the highest percentage of people who walk to work.	Alaska
_____ is the term used to describe a score of three under par in Golf.	Albatross
_____ was used to power the engines of the starship Enterprise in the Star Trek television series.	Antimatter
_____ is a fear of spiders.	Arachnophobia
Dr. John S. Pemberton invented Coca-Cola in 1886 in the city of _____.	Atlanta
The _____ are the present-day name of the land that Columbus called "San Salvador" in 1492.	Bahamas
_____ is the world's tallest grass.	Bamboo
_____ is the most common cause of sport-related eye injuries in the U.S.	Baseball
The _____ is the name of the ship on which Charles Darwin made his famous scientific voyage.	Beagle
The _____ is the national animal of Canada.	beaver
_____ was the first trademarked product.	Beer
The study of plants is known as _____.	botany
A _____ is the term for a newly born whale.	calf
Dominos were invented in the country of _____.	China
_____ is the national sport of England.	Cricket
One of the earliest writing systems was invented by the Sumerians and was known as _____.	cuneiform
_____ is the oldest inhabited city in the world.	Damascus
The car that could travel through time in the film "Back to the Future" was a _____.	De Lorean
_____ is the only word in the English language ending in "mt".	Dreamt
_____ is the name of the ship that Dr. Dolittle and his friends sailed on in the 1967 film Dr. Dolittle.	Flounder
A _____ is what you call a village without a church.	hamlet
_____ is the kind of poison that Socrates took at his execution.	Hemlock
The first atomic bomb exploded in Japan in the city of _____.	Hiroshima
_____ was Time Magazine's "Man of the Year" in 1938.	Hitler
_____ was the first of H.J. Heinz's "57 varieties".	Horseradish
The country of _____ consumes the most Coca-Cola per person.	Iceland
The country of _____ produces two thirds of the world's vanilla.	Madagascar
_____ is the name for the molten rock that erupts from a volcano (forming lava when it cools).	Magma
_____ was the first capital of ancient Egypt.	Memphis
_____ is the only metal that is liquid at room temperature.	Mercury
The planet _____ has surface winds that have been measured at 1500 mph—the strongest in the solar system.	Neptune
_____ is the largest country in Central America.	Nicaragua
The country of _____ consumes more spicy Mexican food than any other European nation.	Norway
The _____ (a type of bird) has eyes that are bigger than its brain.	ostrich
_____ is the alternative common name for a Black Leopard.	Panther
_____ is the term for a group of owls.	Parliament
_____ is the name of the constellation that looks like a flying horse.	Pegasus
_____ was the first US consumer product sold in the Soviet Union.	Pepsi
_____ is the country that has the world's highest railway.	Peru
The first U.S. zoo was built in the city of _____.	Philadelphia
The _____ is the smallest member of the flute family.	piccolo
_____ is the only English word with a completely different meaning when the first letter is capitalized.	Polish
In Roman mythology, _____ is referred to as the "God of the Sea".	Poseidon
_____ is the longest English word without the normal vowels ("a" "e" "i" "o" or "u").	Rhythms
_____ was the first city in the world to have a population of more than 1 million.	Rome
_____ is the only rock regularly eaten by humans.	Salt
There are _____ periods on the periodic table of elements.	seven
_____ is the name of the brightest star in the sky excluding the sun.	Sirius
The country of _____ gave Florida to the U.S. in 1891.	Spain
A _____ has ten tentacles.	squid
_____ was the trade that Greek philosopher Socrates was originally trained for.	Stonecutting
_____ is converted to alcohol during brewing.	Sugar
_____ was the name of the horse that Teddy Roosevelt rode in the Battle of San Juan Hill during the Spanish-American War.	Texas
Angel Falls is located in the country of _____.	Venezuela
_____ is the only planet in the solar system to rotate clockwise.	Venus
The first peanuts in the US were grown in the state of _____.	Virginia
_____ was the first state in the US to allow women to vote.	Wyoming
_____ was the company to first offer a mouse on a commercially available computer.	Xerox
_____ was the last name of the first democratically-elected president of Russia.	Yeltsin
_____ is the financial center and main city of Switzerland.	Zurich

Appendix B. Remote associate triads used in experiment 4:

Triad	Solution
Shock Shave Taste	After
Off Military First	Base
Cry Front Ship	Battle
Bottom Curve Hop	Bell
Control Place Rate	Birth
Cherry Time Smell	Blossom
Child Scan Wash	Brain
Cast Side Jump	Broad
Nose Stone Bear	Brown
Stick Light Birthday	Candle
Wise Work Tower	Clock
Sandwich Golf Foot	Club
Lounge Hour Napkin	Cocktail
Break Bean Cake	Coffee
Sore Shoulder Sweat	Cold
Grass King Meat	Crab
End Line Lock	Dead
Self Attorney Spending	Defense
Back Step Screen	Door
Artist Hatch Route	Escape
Shadow Chart Drop	Eye
Way Ground Weather	Fair
Teeth Arrest Start	False
Land Hand House	Farm
Bump Egg Step	Goose
House Thumb Pepper	Green
Roll Bean Fish	Jelly
Jump Kill Bliss	Joy
Note Chain Master	Key
Reading Service Stick	Lip
Hungry Order Belt	Money
Thread Pine Pain	Needle
Pea Shell Chest	Nut
Blank White Lines	Paper
Fork Dark Man	Pitch
Fence Card Master	Post
Over Plant Horse	Power
Line Fruit Drunk	Punch
Zone Still Noise	Quiet
Horse Human Drag	Race
Home Arm Room	Rest
Test Runner Map	Road
Pet Bottom Garden	Rock
Mate Shoes Total	Running
Oil Bar Tuna	Salad
Lick Sprinkle Mines	Salt
Baby Spring Cap	Shower
Dive Light Rocket	Sky
Iron Shovel Engine	Steam
Pile Market Room	Stock
Forward Flush Razor	Straight
Wet Law Business	Suit
Man Glue Star	Super
Tooth Potato Heart	Sweet
Illness Bus Computer	Terminal
Stop Petty Sneak	Thief
Rope Truck Line	Tow
Mouse Bear Sand	Trap
Pure Blue Fall	Water
Shopping Washer Picture	Window