CHAPTER FIVE

How Retrieval Attempts Affect Learning: A Review and Synthesis

Nate Kornell[*,1] **and Kalif E. Vaughn**[§]
[*]Williams College, Williamstown, MA, United States
[§]Northern Kentucky University, Highland Heights, KY, United States
[1]Corresponding author: E-mail: nkornell@gmail.com

Contents

Psychology of Learning and Motivation, Volume 65
ISSN 0079-7421
http://dx.doi.org/10.1016/bs.plm.2016.03.003

Psychology of Learning and Motivation, First Edition, 2016, 183–215

Abstract

Attempting to recall information from memory (ie, retrieval practice) has been shown to enhance learning across a wide variety of materials, learners, and experimental conditions. We examine the moderating effects of what is arguably the most fundamental distinction to be made about retrieval: whether a retrieval attempt results in success or failure. After reviewing research on this topic, we conclude that retrieval practice is beneficial even when the retrieval attempt is unsuccessful. This finding appears to hold true in a variety of laboratory and real-world contexts and applies to learners across the lifespan. Based on these findings we outline a two-stage model in which learning from retrieval involves (1) a retrieval attempt and then (2) processing the answer. We then turn to a second issue: Does retrieval success even matter for learning? Recent findings suggest that retrieval failure followed by feedback leads to the same amount of learning as retrieval success. In light of these findings, we propose that separate mechanisms are not needed to explain the effect of retrieval success and retrieval failure on learning. We then review existing theories of retrieval and comment on their compatibility with extant data, and end with theoretical conclusions for researchers as well as practical advice for learners and teachers.

1. INTRODUCTION

Hundreds of studies dating back to the late 1800s have reported the mnemonic benefits of retrieval practice (see Roediger & Karpicke, 2006b; for a review). Retrieval practice, or attempting to recall a piece of information from memory, typically produces more learning than not studying, and more impressively, it also produces more learning than restudying (see Roediger & Butler, 2011).

The remainder of this chapter reviews the effects of retrieval success and retrieval failure. Based on a literature review we attempt to answer two main questions. Question 1: Does the act of attempting to retrieve a memory enhance learning even when one does not think of the correct answer (assuming the correct answer is then provided)? Question 2: Is it necessary to propose separate explanations for the effects of retrieval success and the effects of retrieval failure—does retrieval success produce larger effects—or can one set of mechanisms explain both? These questions can be visualized in Fig. 1, which presents seven types of retrieval conditions. In terms of Fig. 1, Question 1 can be asked in three ways: Which is a more effective way to learn, condition 7 or condition 1? What about condition 7 versus condition 2? And how about condition 7 versus condition 3? Question 2, in Fig. 1's terms, boils down to one comparison: If retrieval success versus

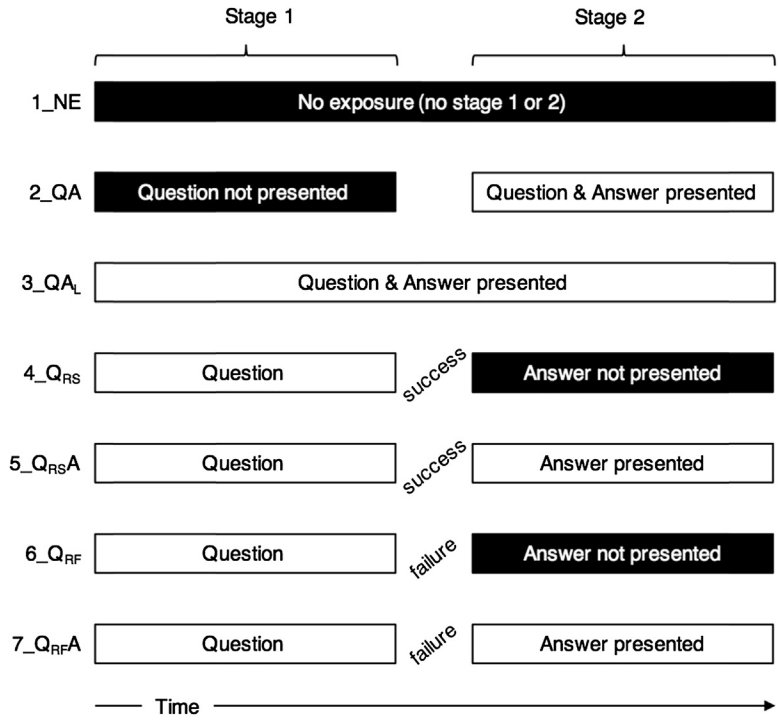

Figure 1 Seven types of trial instantiated in research on retrieval. The left and right halves of the figure represent Stage 1 and 2, respectively, of the two-stage framework. *Black rectangles* represent times when no external stimulus is shown (although in one case the answer has come to mind anyway). *White rectangles* represent times when a stimulus (either a question, answer, or a question/answer pair) are shown externally. Shorthand for each trial type is shown on the left of the figure. In these codes, NE, no exposure; QA, exposure of the question and answer; QA_L, long exposure of the question and answer; Q_{RS}, presentation of a question followed by retrieval success; Q_{RF}, presentation of a question followed by retrieval failure. In these last two cases, an A tacked on to the end means that after the retrieval attempt the answer was shown.

failure is experimentally manipulated, which is a more effective way to learn, condition 5 or 7?

We begin by discussing research in which retrieval success does seem to be an important moderating factor, but then we discuss limitations with these studies. Then we review research showing that retrieval attempts enhance learning even when they are unsuccessful. Next, we demonstrate that when retrieval success is manipulated, participants learn just as much following a failure as following a success. We also discuss the strengths and weaknesses of existing theories of retrieval practice. We end by discussing practical implications for educators.

2. THREE KINDS OF EVIDENCE

In this section we review three kinds of evidence that can be used to answer our first question, namely, do unsuccessful retrieval attempts enhance learning. To foreshadow, we conclude that a convincing answer can be obtained only by looking at the third type of evidence.

2.1 Retrieval Difficulty

By definition, as retrieval difficulty increases, retrieval success decreases. If retrieval success were necessary for learning then, all else being equal, one would expect an increase in retrieval difficulty to produce a corresponding decrease in learning, because more retrieval difficulty implies less retrieval success. For example, students who are given a relatively difficult set of questions and fail to retrieve most of the answers should benefit less from the experience than students who are given easier questions and experience a lot of retrieval success.

This pattern of data can occur when feedback is not given after testing. In Fig. 1, in a comparison of conditions 2_QA and 4_Q_{RS}, the latter might produce more long term learning (eg, Roediger & Karpicke, 2006a). But in a comparison of conditions 2_QA and 6_Q_{RF}, the former will almost certainly produce more learning because in the latter, the participant would never be exposed to the correct answers. However, we believe that condition 6_Q_{RF} is exceedingly rare in education: When people cannot think of the answer to the question, they try to figure it out; when they cannot think of the word on the back of a flashcard, they turn the card over; and when a teacher asks a question and the students get it wrong, she tells them the correct answer. The questions we focus on concern comparisons of conditions 2_QA, 3_QA_L, 5_Q_{RS}A, and 7_Q_{RF}A in Fig. 1 (Research by Pashler, Cepeda, Rohrer, & Wixted, 2005 suggests that conditions 4_Q_{RS} and 5_Q_{RS}A result in the same amount of learning, so what we say about condition 5 applies to condition 4). In other words, we focus on situations when participants who do not answer correctly are told the correct answer.

When feedback is given following retrieval attempts, the data do not show that easier retrieval is more effective. They show the opposite pattern: Retrieval practice is more effective when the learning conditions promote increased retrieval difficulty. Such experimental conditions include objectively more difficult versus easier practice tests (eg, free recall versus multiple

choice; Duchastel, 1981); interleaved versus blocked practice (ie, mixing various problem types in one practice session versus massing practice of one problem type; Rohrer & Taylor, 2007); spaced versus massed practiced (ie, distributing practice across a longer time period versus cramming practice into a shorter time frame; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006); or a longer versus shorter practice lag (ie, a larger stack of to-be-learned information makes retrieving any one item more difficult; Kornell, 2009; Pashler, Zarow, & Triplett, 2003). The common thread in these studies is that creating difficulty during learning impedes initial performance, but enhances final test performance. Manipulations that increase initial acquisition difficulty but enhance delayed memory performance have been referred to as *desirable difficulties* (see Bjork, 1994; Bjork & Bjork, 2011). By demonstrating situations in which more retrieval success is associated with less learning, these findings cast a shadow over the hypothesis that retrieval success is integral to the benefit of testing.

Importantly, these findings do not prove that unsuccessful retrieval attempts enhance learning, because comparisons are being made between different learning strategies that are associated with different levels of retrieval success, meaning that other factors are influencing the results. For instance, the benefit of spaced practice may far outweigh the lower retrieval success it affords (resulting in a positive spacing effect even if retrieval failures followed by feedback produce no learning). Given that comparing the effect of two or more learning manipulations that affect retrieval success is necessarily confounded, we move on to discuss the role of item difficulty on retrieval success.

2.2 Item Difficulty

Across any set of materials, certain items are easier to learn and recall than others. Difficulty can vary both objectively (ie, certain items are difficult to learn for most people) and idiosyncratically (ie, certain items are difficult for a particular learner). Retrieval success is lower for items that are more difficult. Thus, one might expect less learning to occur for difficult versus easy items due to the lower rate of retrieval success. However, such a comparison is confounded with item difficulty: Long-term benefits of retrieval success (or failure) could be due to differences in the learnability of the easy versus difficult items, and not based on retrieval success itself. Given that comparing different sets of items is necessarily confounded, we do not see this as a promising research approach.

2.3 Experimental Control of Retrieval Success

Confounding variables can be avoided when the experimenter controls retrieval success and failure. Two categories of studies have exerted such control. One is studies in which retrieval success is held constant (ie, retrieval success never happens). These studies have investigated our first question, namely whether retrieval attempts enhance learning even when they are unsuccessful. The other is studies in which retrieval success is manipulated experimentally. These studies have investigated our second main question: Does retrieval success improve learning more than retrieval failure? Studies that have directly controlled retrieval success are free of confounding variables and thus provide the most convincing form of evidence pertaining to the benefits of retrieval success. (As we point out in the General Discussion, however, these studies might have less practical importance than studies that manipulate learning strategies and item difficulty, which we described in Sections 2.1 and 2.2, because in real life, unconfounded control of retrieval success is rare.)

3. WHY RETRIEVAL SUCCESS MIGHT MATTER

Before discussing studies that have manipulated retrieval success, we briefly review two kinds of evidence involving retrieval success. At first glance, both kinds of evidence suggest that retrieval success might matter, but on closer inspection neither provides strong evidence either way. First, items that have been successfully retrieved more times tend to be learned better than items that have been successfully retrieved fewer times. Several studies have shown that setting the criterion level (ie, the number of times an item must be recalled during practice before retrieval practice on the item ceases) higher increases learning (Pyc & Rawson, 2009; Rawson & Dunlosky, 2011; Vaughn & Rawson, 2011, 2014). These studies suggest that retrieval success is a good thing, but they compare more retrieval practice to less, not more success to an equal amount of practice with less success. In other words, they do not have the appropriate control condition (eg, a condition in which items received the same amount of test practice without retrieval success) to ascertain the extent to which retrieval success per se was responsible for the better final test outcomes as opposed to simply more versus less test practice in general.

Another way to examine retrieval success is to use contingency analyses. Research has shown that if an item was correctly recalled on an initial test, there is a high probability that it will also be correctly recalled on a delayed

test (similarly, if the item was not correctly recalled initially, there is a low probability that it will be correctly recalled on a delayed test; eg, Kahana, 2002). These findings are also not satisfying for our purposes. Items that are recalled on an initial test are, on average, easier than items that are not recalled. Thus, differences in item difficulty could explain differences in final test performance: Easy items will be recalled initially and on the final test because they are easy, whereas difficult items will not because they are difficult. In other words, contingency analyses are not useful for our purposes because of item selection effects (see Kornell, Hays, & Bjork, 2009; Pashler et al., 2003). Next, we review studies in which retrieval success is under experimenter control.

4. EVIDENCE THAT UNSUCCESSFUL RETRIEVAL IMPROVES MEMORY

We have now outlined a litany of evidence consistent with the idea that retrieval failure might enhance learning. However, none of that research was convincing and one might worry that the opposite is true. Unsuccessful retrieval attempts involve making an error, and perhaps it would be better to restudy without first making a memory error than to make the error and then restudy. Traditionally, commission errors (ie, giving the wrong answer) have been treated as more worrisome than omission errors (ie, failing to give any answer)—partly because in most research with animals, commissions are the only kind of error. B. F. Skinner viewed learning as a constant reinforcement process, and an ideal learning environment was one that involved a constant progression of difficulty at a pace that minimized error production (eg, Skinner, 1968). This so-called "errorless learning" was thought to produce the best outcomes, as it was typically believed that producing an error causes the learner to reinforce that error, impeding subsequent learning (eg, Guthrie, 1952; Terrace, 1963). However, modern research has not always accorded with the idea of errorless learning; for example, the more confident one is in an error, the *more* likely one is to correct that error on a subsequent test (see Butler, Fazio, & Marsh, 2011; Butterfield & Metcalfe, 2001). The focus of this section is to summarize research that has overturned the assumption that memory errors are detrimental to learning.

4.1 Test-Potentiated Learning

Izawa provided the first systematic investigations of how retrieval failures can improve learning. In a typical study, Izawa (1970) provided learners with a

series of test trials (T) followed by a study opportunity (S). She manipulated how many test trials occurred before the eventual study trial occurred. For instance, Izawa (1970) examined five separate learning conditions: ST, STT, STTT, STTTT, and STTTTT. In each condition the cycles repeated (eg, STSTST, STTSTT, and so on) for a total of 25 trials. Thus, the ST condition received 13 study trials and 12 test trials, and the STTTTT condition received 5 study trials and 20 test trials. This procedure is outlined in the left panel of Fig. 2.

In this study, and other similar studies conducted by Izawa (eg, 1969), Izawa plotted how often participants made errors during the test trial that followed a restudy trial. As the experiment progressed, the error production rate decreased across all conditions; however, a restudy trial caused the greatest reduction in errors when preceded by more (eg, STTTTT) versus fewer (eg, ST) tests. Izawa referred to this effect as "test-potentiation," because retrieval failures seemed to increase, or potentiate, the amount of learning that occurred on subsequent study trials. These findings suggest that unsuccessful retrieval attempts can actually improve learning.

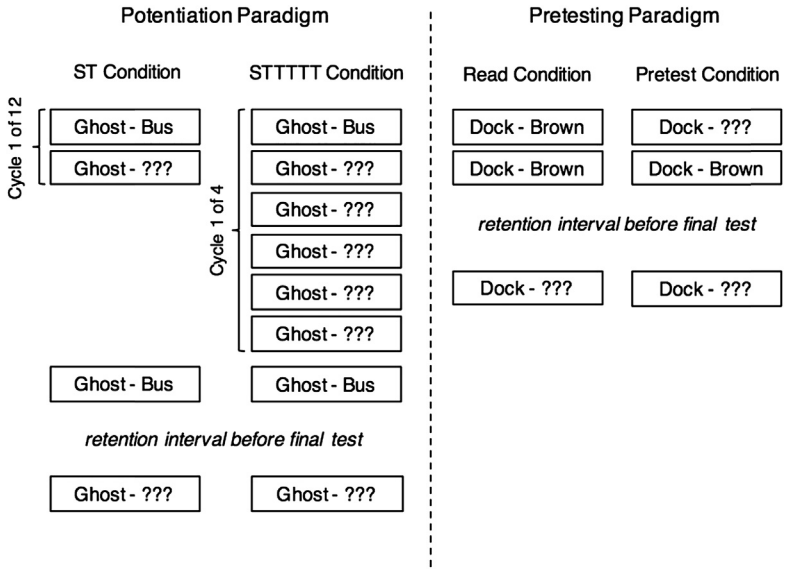

Figure 2 Paradigms used to study test-potentiated learning (left) and pretesting (right). In the example of on the left, both conditions include 24 trials followed by a final feedback trial, for a total of 25 trials, during learning. All four of the depicted conditions involve learning multiple pairs, but only one of the pairs is shown. The ST Condition and Read Condition both serve as control conditions in which there is less testing than in the comparison conditions (STTTTT and Pretest).

Arnold and McDermott (2013) replicated Izawa's findings using a similar paradigm but with real-world materials. While Izawa's (1970) participants studied pairs consisting of a three-letter nonword paired with a noun, Arnold and McDermott (2013) had participants study 25 Russian—English word pairs (eg, medved—bear) during the study phase. During a test phase, participants either received one test trial per item (ST) or five test trials per item (STTTTT). These cycles repeated until participants had received 20 study and test trials, with 4 study trials and 16 test trials in the STTTTT condition and 10 study trials and 10 test trials in the ST condition. Arnold and McDermott (2013) found that restudy trials were more beneficial when more (ie, five) versus fewer (ie, one) tests preceded the restudy trial. (From a practical standpoint, it is worth mentioning that at the end of learning, participants had learned more in the ST condition than the STTTTT condition in the studies by both Izawa (1969, 1970) and Arnold and McDermott (2013), so test potentiation was not an efficient use of time overall. The important point here, though, is that an individual study trial was more beneficial if it was preceded by more unsuccessful retrieval attempts.)

4.2 Pretesting Procedures

Kornell et al. (2009) used a different paradigm to investigate the benefits of unsuccessful retrieval attempts (see the right panel of Fig. 2). In order to minimize item-selection effects, Kornell et al. used a pretesting procedure that insured participants would hardly ever answer correctly prior to being given feedback. Their learning materials were weakly related paired associates and there was no initial study phase. The first encounter participants had with a word pair was a test (eg, whale—???) followed by feedback (eg, whale—mammal). The failure rate on this initial test was around 95% across Experiments 3—6 (Kornell et al., 2009) and to insure retrieval failure, any items participants did correctly recall were excluded from the data analysis. Items in the read-only condition received either 5 s of study (Experiment 3) or 13 s of study (to control for time on task; see Experiments 4—6). In terms of Fig. 1, Experiment 3 compared conditions 2_QA and $7Q_{RF}A$; Experiments 4—6 compared conditions 3_QA_L and $7\_Q_{RF}A$.

Across all experiments, Kornell et al. found that final test performance favored items that had been pretested versus read. This is a surprising finding given that the pretest trials resulted in the production of a large number of memory errors. (It is especially surprising when one considers that any items correctly retrieved on the initial test were discarded from subsequent

analyses). Kornell et al. replicated this pretesting effect with new materials (ie, fictional questions with no correct answer in Experiments 1 and 2), across a delay (about 38 h in Experiment 5), and using a between-participants manipulation (Experiment 6).

Other studies have also shown the benefits of making errors on subsequent learning. For instance, Potts and Shanks (2014) presented participants with obscure English words (eg, frampold) and had them read, generate, or choose the correct meaning from two possible choices (eg, quarrelsome). Results showed that generating and attempting to choose the correct answer, even when initial performance was near chance, improved final performance compared to a read-only condition. Thus, as in Kornell et al. (2009), error production during initial learning was associated with enhanced memory on a final test.

Although the pretesting effect is intriguing and spawned additional research on the possible benefits of unsuccessful retrieval attempts (eg, Grimaldi & Karpicke, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Richland, Kornell, & Kao, 2009; Vaughn & Rawson, 2012), there is a potential limitation with these studies: Because participants were not given an initial study phase during the experiments, they were essentially guessing when attempting to answer, and their retrieval attempt did not involve *episodic* retrieval of a word pair that they had previously studied. Other studies have implemented the pretesting procedure in more realistic situations that do not involve guessing and have obtained similar results (as we explain in Section 6, Moderators of the Testing Effect).

5. THE TWO-STAGE FRAMEWORK

Before we describe potential moderators of the pretesting effect, we pause briefly to fit the findings from test-potentiation procedures (eg, Arnold & McDermott, 2013; Izawa, 1970) and pretesting procedures (eg, Kornell et al., 2009) into a two-stage framework.

Kornell, Klein, and Rawson (2015) have proposed a framework in which learning from retrieval is conceptualized in two stages: (1) the learner attempts to retrieve the correct answer (ie, a retrieval attempt occurs), and (2) the correct answer becomes available (ie, the learner thinks of the answer or feedback is given). This framework is illustrated in Fig. 3, and it can also be seen in the left (Stage 1) and right (Stage 2) columns in Fig. 1. The two-stage framework is founded on the uncontroversial observation that there is a

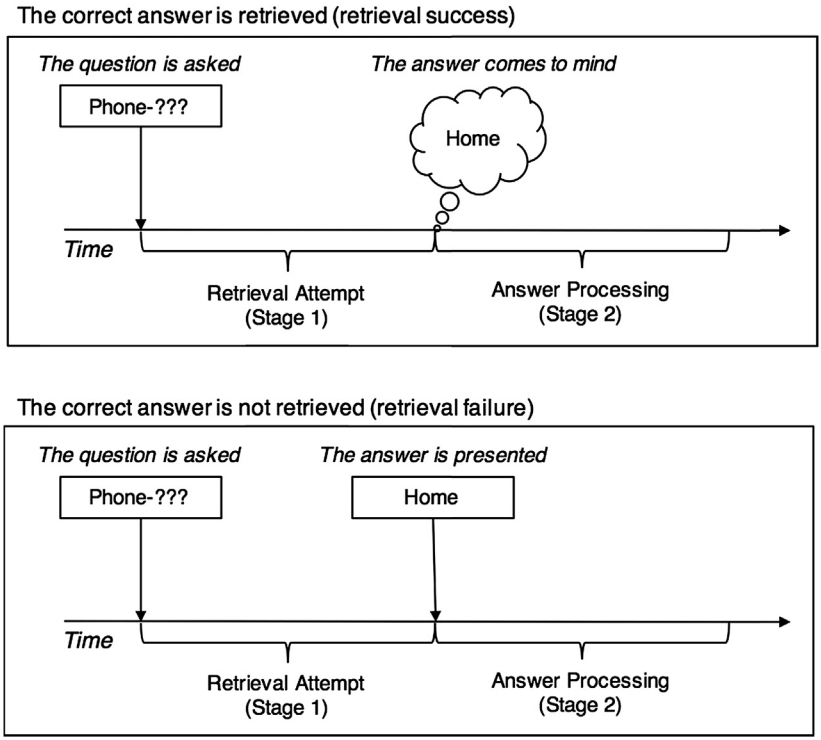

Figure 3 Schematic highlighting both stages of the two-stage framework: Stage 1 refers to the initial retrieval attempt, and Stage 2 refers to the post-processing that occurs during feedback. In the top panel retrieval is successful. In the bottom panel, the answer is not retrieved but the participant is given feedback.

naturally occurring, nonarbitrary division in the cognitive processes that underlie learning from retrieval: there is pre-answer processing and post-answer processing. Some might disagree with calling Stage 2 a part of the retrieval process, because the answer has already been retrieved. So let us be clear: This framework does not describe the process of retrieval (which ends at the end of Stage 1), it describes the process of *learning* from retrieval. Stage 2 is clearly part of this learning process: As the results we reviewed in the previous section showed, Stage 2 became more effective as a result of Stage 1 even though Stage 1 did not actually involve learning the answer.

The two-stage framework does not posit mechanism to explain the benefits of retrieval. It is not a theory and it makes no predictions. But thinking in terms of the two stages can help inform theory. As a case in point, through the lens of the two-stage framework one can see that the same mechanism

might be responsible for test-potentiation effects and pretesting effects (Kornell et al., 2015). Although the paradigms are slightly different (see Fig. 2), in both lines of research there is a retrieval attempt (ie, Stage 1) and then processing of the answer (ie, Stage 2). Moreover, in both cases, the data can be explained based on the idea that more Stage 1 processing leads to more learning in Stage 2. To explain why this is the case would require a theory. Although we return to theory at the end of this chapter, for now we just note that it is parsimonious to assume, until data prove otherwise, that unsuccessful retrieval attempts in these two paradigms enhance memory via the same basic mechanism: Engaging in Stage 1 processing makes the processing that happens in Stage 2 more effective. Of course, we cannot rule out the possibility that test-potentiation and pretesting operate via different mechanisms, but more data would be needed to support that claim. (To foreshadow, later in the chapter we make a similar, but stronger, claim. We believe the same mechanism that explains potentiation and pretesting might also explain the effect of successful retrieval attempts.)

6. MODERATORS OF THE PRETESTING EFFECT

There has been an uptick in research on pretesting since 2009. Next we review this research, which has examined the effect of a variety of moderating variables (eg, the timing of feedback, the materials being learned, and the age of the participants) on pretesting.

6.1 Feedback Timing

In the pretesting studies described thus far, feedback was given immediately following the test trial. But what if a learner fails to retrieve the correct answer, and then does not receive immediate feedback? The consensus seems to be that, as long as feedback is provided at some point, unsuccessful retrieval attempts (or maybe they should be called guesses) do not hurt future learning of the correct response. However, they do not necessarily help either.

Grimaldi and Karpicke (2012) had learners complete either a pretest trial with immediate feedback or delayed feedback (in the latter condition a block of pretest trials was followed by a block of feedback trials). They also included a no-pretest control condition. Performance in the pretest/immediate feedback condition exceeded performance in the other two

conditions, which were about the same. Therefore, when feedback was delayed, pretesting did not seem to enhance subsequent learning (for replications, see also Hays, Kornell, & Bjork, 2013; Vaughn & Rawson, 2012).

6.2 Trivia Questions

Paired associate studies suggest that immediate feedback is necessary for pretesting to produce a benefit, but the same does not seem to be true for more complex materials. Kornell (2014) investigated the benefits of attempting to answer memorable but largely unknown trivia questions (eg, Q: What was the first state to allow women to vote? A: Wyoming). Importantly, Kornell investigated the benefit of guessing the answer to trivia questions both with immediate feedback and delayed feedback. Experiment 2 compared a condition in which there was a delay of around 6 min between when the guess occurred and when feedback was provided to a condition in which the feedback was immediate. These conditions produced similar performance, and both were better than a study-only control conditions. These results suggest that delaying the feedback following a guess trial was not detrimental. In Experiment 3, Kornell replicated this result with a much longer delay between the guess trial and feedback trial. Experiment 3 had three sessions. During Session 1, half of the items were tested three times and the other half of the items received no exposure. After a 24-h delay, participants completed Session 2. During Session 2, all items received a study trial (ie, feedback). After another 24-h delay, participants completed Session 3, which involved taking a final test on all the items. If pretesting only enhances learning when feedback is immediate, then there should be little or no benefit of pretesting when feedback is delayed by 24-h. In reality, performance was greater when items were tested initially, suggesting that guessing was beneficial even when feedback was delayed by 24-h. Even more impressive is the fact that these results persisted across a 24-h retention interval, suggesting that attempting to answer questions without a prior study phase can benefit learning on both immediate and delayed tests.

Why would attempting to answer trivia questions benefit learning even when feedback was delayed by up to 24-h, whereas attempting to guess the answer to cue-target associations require immediate feedback in order to benefit learning? One possibility is that cue—target associations evoke a more haphazard form of guessing, whereas trivia questions tap into a rich semantic network of possible answers. For instance, participants presented with "whale—????" know that there is no actual solution to this problem and will likely base their guess on whatever associate springs to mind. In

contrast, participants presented with a question like: "What was the first state to allow women to vote?" will think there is an actual solution to this question, and will thus begin searching their semantic memory for the correct answer. They may activate various nodes in their semantic network such as the United States, suffrage movements, progressive politics, mavrickiness, and so forth. There is less semantic information to activate when presented with a cue–target pair such as "whale–????" and, because the information does not form a rich web of semantic activation, the priming of the activated information may fade quickly from memory. Consistent with this idea, researchers have found that pretesting effects that occur when items are related (eg, tide–beach) disappear when the items are not related even if feedback is immediate (eg, stem–candy; Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012). It appears that activating meaningful questions in memory enhances subsequent learning even if feedback is delayed. They also suggest that unsuccessful retrieval attempts can enhance learning even when the question is meaningful and there is an actual correct answer, not only when participants are guessing.

6.3 Scholastic Materials

Unsuccessful test practice seems to help for both cue–target pairs and general knowledge questions (as well as for nonsensical information such as the nonwords used by Izawa), but what about for more educationally relevant material? Richland et al. (2009) had participants read an essay on vision. Embedded within the essay were key concepts and definitions pertaining to a vision disorder (cerebral achromatopsia), from which 10 test questions were created. For participants in the pretest condition, half of the items received a pretest prior to the study phase (ie, prior to reading the passage). Pretesting lasted for 2 min, during which time participants were told to answer all five questions (even if they had no idea what the correct answer was). Afterwards, pretest participants were given 8 min to read the passage. Participants in the extended study condition were given 10 min to study, so that total time on task was equated between the extended study and pretest conditions. On an immediate final test, performance was best for items that had been pretested. Performance was nearly equivalent for the untested items in the pretest condition and the items in the extended study condition (see Experiment 1). Richland et al. (2009) went on to replicate this pretesting effect in four additional experiments, controlling for possible encoding differences (eg, by bolding and/or italicizing important words in the passage for participants in the extended study condition) and by providing evidence

that attempting to answer the pretest questions was the specific mechanism that boosted subsequent performance (instead of just providing a useful framework for reading the passage or providing some kind of deeper processing; see Experiment 5). Overall, the results from Richland et al. (2009) provide evidence that pretesting enhances memory for educationally relevant material more than spending an equivalent amount of time studying. These results also replicated the finding that when the learning materials are semantically rich, pretests can be beneficial even when the feedback is delayed by a few minutes.

A study by Kapur and Bielaczyc (2012) was an even better approximation of a real educational experience. They demonstrated the benefits of pretesting on the learning of mathematics concepts in a classroom setting. Math students in Singapore were assigned to one of two conditions. The *productive failure* group attempted to solve math problems without any explicit instruction, while the *direct instruction* group practiced solving math problems with instructional supervision and feedback. The former condition resulted in a large number of math errors, which were eventually corrected during a final session that involved a teacher explaining how to correctly solve the problems; the latter condition resulted in students making fewer errors and having them corrected more quickly. On a posttest, the productive failure group outperformed the direct instruction group, despite having made a larger number of errors during practice. These results replicate the finding that pretesting improves learning, and extends those results by showing that pretesting helps with complex math materials in a real-world classroom.

6.4 Older Adults

Although pretesting seems to enhance learning, every study we have reviewed so far used younger adults as participants. Error production seems to be more problematic in older adults, and particularly in older adults with memory impairments (eg, Alzheimer's disease). For instance, Glisky, Schacter, and Tulving (1986) developed a method of vanishing cues in order to minimize error production during practice for participants with brain injuries. At first, fragment hints are provided with as many letters as needed to facilitate recall success. Then, once a correct response is given, letters are removed from the fragment hint until the hint is no longer needed at all. Glisky et al., found that for participants with brain injuries, the vanishing cue condition (which minimized error production and constituted errorless learning) led to better learning outcomes than a typical paradigm without hints (which involved more errors and constituted errorful learning).

Squires, Hunkin, and Parkin (1997) showed that participants with amnesia who were presented with novel associations (with unfamiliar word pairs, eg, card—spade) learned more in an errorless condition than an errorful condition, and even provided some evidence that the benefit persisted on a delayed test (see their Experiment 2, although that effect was not found in all of their experiments). Wilson, Baddeley, Evans, and Shiel (1994) reported a series of experiments that showed benefits of errorless learning across a wide variety of learning stimuli for participants with amnesia (eg, learning the names of objects and people, learning general knowledge information, and also learning how to program an electronic device). Although researchers do not always find a clear benefit of errorless learning (eg, Haslam, Gilroy, Black, & Beesley, 2006), the overall pattern of data suggests that error production may be detrimental for older adults and other people with memory impairments. It does not appear to be detrimental for younger learners without memory impairments (for a review, see Clare & Jones, 2008).

Although errors may be detrimental for memory-impaired older adults, there is a potential caveat with this conclusion. Cyr and Anderson (2014) point out that the difference between older and younger adults is confounded with procedural differences: Research with younger adults tends to rely on conceptual information (eg, cue—target relationships) whereas research with older adults tends to use more lexical or nonconceptual information (eg, a list of words). It is not clear, therefore, why errorless learning works in one case and not the other—the longstanding assumption has been that the key difference is the memory abilities of the participants, but Cyr and Anderson postulated that procedural difference might actually be the key, and they predicted that errors may be beneficial to the learning process for conceptual information regardless of age (ie, for both young and old adults). To examine this issue, Cyr and Anderson (2014) manipulated errorless and errorful learning for both younger and older adults. Participants in both age groups learned either words associated with a semantic category (eg, a type of fruit), which were conceptual, or word stems (eg, fl_____), which were not conceptual. On a final test after a 10-min delay, performance for conceptual information (recalling the categorized words) was highest in the errorful learning condition. Conversely, performance for the lexical information (recalling stem words) was highest in the errorless learning condition. Crucially, these effects occurred for both younger and older adults. These results lend credence to the idea that error production is not inherently beneficial or detrimental for any age group, nor does it appear

to depend on age; rather, the effect of errors may depend on other factors such as the type of materials used and the type of processing afforded by them (for similar results, see Haslam, et al., 2006; McGillivray & Castel, n.d.).

6.5 Metacognitive Awareness

Error production seems to benefit learning, but are participants aware of these benefits? Huelser and Metcalfe (2012) addressed this issue by having participants make metacognitive judgments after learning and being tested on either weakly related or unrelated word pairs. Ninety cue—target pairs were divided into three learning conditions during the practice phase: read short, read long, and error generation (ie, pretesting). On a read short trial, both the cue and target were presented for 5 s. During a read long trial, both the cue and target were presented for 10 s. During an error generation trial, the cue was presented on the screen for 5 s (allowing participants to type in their answer), followed by 5 s of cue—target feedback (in Fig. 1, these conditions correspond to conditions 2_QA, 3_QA_L, and 7_$Q_{RF}A$, respectively). When the learning materials were pairs of related words, pretesting produced the best final test performance in both their experiments (as we mentioned earlier, no difference between conditions emerged for unrelated pairs in either of their experiments). After the final test, participants made metacognitive judgments about the effectiveness of the three trial types. Participants were asked to rank the trial types from best (ie, what they thought helped them learn the word pairs the best) to worst (ie, what they thought helped them learn the word pairs the least). Participants ranked the read long trial as the most effective for both related and unrelated materials, whereas pretesting received the lowest utility rating for both sets of materials. These ratings clearly represented a metacognitive error because actual accuracy was highest, not lowest, in the pretesting condition, at least with related words. In short, Huelser and Metcalfe's participants viewed errors as detrimental even when they were actually beneficial.

7. DOES RETRIEVAL SUCCESS EVEN MATTER?

We began this chapter by posing two questions. We have now reviewed evidence that seems to provide an answer to the first: Attempting to retrieve a memory enhances subsequent learning even if the attempt is unsuccessful. This conclusion leads us to our second question: Does retrieval success even matter?

It seems reasonable to think that retrieving information from one's own memory would be more beneficial than receiving external feedback. For example, arriving at an answer via one's own semantic network might do more to strengthen one's memory than being directed to the answer externally. This hypothesis fits with the idea that retrieval is beneficial because it strengthens so-called retrieval routes (eg, Bjork, 1975). But it is important to test this hypothesis, because if successful and unsuccessful retrieval lead to the same amount of learning, it may be parsimonious to propose that they should be explained by the same underlying mechanism.

It is difficult to test the hypothesis that successful retrieval is more beneficial than unsuccessful retrieval because of item selection effects. Items correctly recalled on an initial test are easier items for that participant, and comparisons to items not recalled on the initial test are unfair because they conflate item difficulty with retrieval success. To compare retrieval success to retrieval failure requires using random assignment to determine whether or not participants succeed when they make a retrieval attempt. This cannot be done using pretesting or potentiation, because both paradigms look specifically at retrieval failure and do not allow an analysis of retrieval success.

Ideally, in order to directly examine the influence of retrieval success compared to retrieval failure on learning, we would need a paradigm that met two criteria. In addition to random assignment, the paradigm would need to involve an initial study phase, so that subsequent test trials reflected episodic retrieval and not random guessing. Kornell et al. (2015) developed a paradigm that met these criteria. First, Kornell et al. administered an encoding phase during which weakly related word pairs received two copy trials administered in two separate blocks. During a copy trial, participants studied the word pairs (eg, gamble: chance) with instructions to copy the target word on a similar line appearing below the cue–target pair (eg, gamble: ch___e). After copying the words twice, the practice phase began. During a practice trial, the cue word was presented with instructions to retrieve the matching target word (eg, gamble: ch___e). Immediately after the test trial, the experimental manipulation occurred: Half of the items received a fragment trial and half received a copy trial. Fragment trials were designed to facilitate retrieval success and involved presenting the cue word along with a fragment of the target word (eg, gamble: ch___e). Copy trials functioned the same as during initial study. In other words, during the practice phase participants always made a retrieval attempt; then they were either given a hint that allowed them to retrieve or they did not

retrieve but were given feedback. After the practice phase, a final test occurred on which the cue was shown and participants were asked to type in the target. Items that participants retrieved during the practice phase (prior to being shown the fragment or target) were excluded from the analyses because the point of the procedure was to randomly assign items to be retrieved successfully (via fragment trials) or not (via copy trials). Kornell et al. found that the fragment and copy conditions produced almost equivalent performance on the final cued recall test (there was a small but significant advantage for the fragment condition in one of their experiments but this advantage flipped in another). In other words, for items that were not successfully recalled initially during practice, it did not matter whether or not those items went on to be successfully recalled (ie, received a fragment trial) or simply received restudy feedback (ie, received a copy trial).

These findings suggest that the reason retrieval is beneficial is not retrieval per se. Instead, the benefit actually hinges on making a retrieval attempt. Before discussing the implications of this finding, in the next section we present research designed to replicate and extend this finding.

7.1 What About Target Memory?

Kornell et al. (2015) measured learning based on cued-recall. Cued recall tests reflect two types of memory: associative memory linking the cue to the target (eg, the link from *gamble* to *chance*) and target memory (eg, the word *chance*). Prior studies have shown that retrieval improves memory for cues, targets, and their associations (eg, Carpenter, Pashler, & Vul, 2006; Vaughn & Rawson, 2011, 2014). Even though retrieval success did not seem to affect associative memory, it might impact target memory. Notice that in most prior studies investigating retrieval success (whether based on a pretesting procedure or a test-potentiation procedure), the final test was a cued recall test, making it difficult to know the extent to which retrieval influenced target memory. Memory for the word that is being retrieved—the target word—might be especially sensitive to whether retrieval of the target word was successful.

To test this hypothesis, we used the same procedure as Kornell et al. (2015) with one key exception: instead of a final cued-recall test, we used a target recognition test to measure learning. The target recognition test was administered after a two-day delay and consisted of 40 targets and 40 lures being presented one at a time in a random order, along with the question, "Was this one of the words you studied earlier in the experiment?" By

using a final target recognition test, we assessed the extent to which retrieval success versus retrieval failure influenced target memory.

The accuracy of recognition memory was not affected by whether participants' retrieval attempts were ultimately successful (ie, they unsuccessfully attempted retrieval and then completed a fragment trial) or not (ie, they unsuccessfully attempted retrieval and were then shown the target). These data conceptually replicated the results of Kornell et al., (2015) and suggest that as long as a retrieval attempt occurs, retrieval success does not matter, even for target memory.

7.2 Fragments as Feedback

Before discussing the implications of our findings and those of Kornell et al. (2015), we will comment the methodology used in these studies. We wish to reiterate that our feedback manipulation was only used as a tool to evoke retrieval success in the fragment condition. We do not claim it is perfect; seeing fragments and thinking of the answer spontaneously are not identical experiences. We happily concede that if our method is flawed then our conclusions should be treated with suspicion. However, we believe this method does accomplish the goal of manipulating retrieval success, despite its limitations.

Moreover, other researchers have used similar feedback manipulations in an effort to improve learning. For instance, Finn and Metcalfe (2010) provided scaffolding feedback (ie, presenting additional letters of the correct answer until the participant correctly recalled the answer) after participants failed to come up with a correct response to a question. Their results demonstrated more learning from scaffolded feedback than from other forms of feedback. (As we mentioned earlier, similar results were found with amnesic patients; Glisky & Schacter, 1989; Glisky et al., 1986.)

One might wonder why Finn and Metcalfe (2010) found that scaffolding feedback enhanced learning (compared to standard feedback), whereas our results, and those of Kornell et al. (2015), demonstrated that fragment feedback did not enhance learning (compared to standard feedback). One possible explanation lies in a procedural difference: In the present study, a retrieval failure was followed by either standard feedback or one very easy chance to retrieve the answer. In Finn and Metcalfe's (2010) scaffolding procedure, on the other hand, feedback was presented one letter at a time until the participant came up with the answer. There are two reasons why this difference might be important. One is the number of separable retrieval attempts. In our procedure, the fragment was presented once; Finn and

Metcalfe presented many versions of the fragment as letters were added. The other, which might be more important, is that in our procedure it was easy to think of the answer when shown the fragment and in many cases, participants probably could have thought of it with fewer letters. Finn and Metcalfe guaranteed that their participants (usually) retrieved the answer when it was as difficult as possible, that is, as soon as there were enough letters for them to think of the answer, so they never got additional letters that would have made it easier. On the other hand, our results suggest that retrieval success versus retrieval failure followed by feedback should not matter, and the initial retrieval attempt with no target letters was already maximally difficult for both groups. More research is needed to understand why our findings, and similar findings by Kornell et al. (2015), differ from Finn and Metcalfe's (2010) findings.

8. THEORIES OF TEST-ENHANCED LEARNING

So far in this chapter we have discussed three ways retrieval might affect learning: (1) by potentiating subsequent study following a retrieval failure in the potentiation procedure, (2) by doing the same thing in the pretesting procedure, or (3) by enhancing learning directly when retrieval attempts are successful. Although these three situations have different surface features, it is not clear that they are different at a deep level. In fact, we hypothesize that a single set of mechanisms might be responsible for how people learn from retrieval regardless of whether retrieval is successful or not. In terms of the two-stage model, we would say that the retrieval attempt in Stage 1 always potentiates learning when the answer is available in Stage 2, regardless of whether the answer became available via retrieval success or feedback. However, the two-stage model does not posit a mechanism. Thus, next we review existing theories that predict benefits of retrieval. In each case, we examine the theory through the lens of the two-stage framework and discuss each with respect to retrieval success versus retrieval failure.

8.1 New Theory of Disuse

The *new theory of disuse* (for brevity, NTD; eg, Bjork & Bjork, 1992, 2011) discusses memory through two strength properties: retrieval strength and storage strength. Retrieval strength refers to how readily an item can be accessed in memory at the moment, whereas storage strength refers to

how well learned that item is in memory. Retrieval practice enhances both retrieval strength and storage strength, but the degree to which these strengths are enhanced depends on the level of each at the time of the test. For instance, gains in storage strength are largest when retrieval strength is low (which corresponds to more retrieval effort). Conversely, the higher the storage strength, the smaller the gains in retrieval strength (ie, the more well-learned an item is, the harder it is to make additional gains). Once accumulated, storage strength is never lost. Retrieval strength fades over time, but the extent to which it fades depends upon the level of storage strength. A higher level of storage strength minimizes losses in retrieval strength; however, a low level of storage strength means that any gains in retrieval strength will be short-lived without additional learning. In sum, NTD posits that retrieval strength and storage strength dynamically influence the benefit of a learning trial.

Test trials are thought to be more potent than study trials in NTD, but this is simply an assumption of the theory, no mechanism is provided to explain how it happens. However, the theory is relevant to retrieval success versus failure. To be clear, the theory does not make predictions about retrieval success versus failure per se. However, it does make predictions about retrieval ease versus difficulty, which are associated with retrieval success versus failure. Using NTD's terms, when retrieval strength is low, there should be (1) more retrieval failure and (2) more learning. Thus, NTD seems to imply that if anything, retrieval failures might lead to more learning than retrieval successes. For example, in a pretesting procedure, storage strength and retrieval strength are necessarily low (if not nonexistent) because no prior study phase has been given for those items. In contrast, items given a prior study phase would have some level of storage strength and retrieval strength. In this case, NTD might predict that the benefit of retrieval should be greater for unstudied items than for items that have been studied before.

This comparison requires a clarifying comment. Researchers do not typically examine the effects of retrieval in a vacuum. Instead, they typically compare retrieval to restudy. Thus, one might interpret the phrase "benefit of retrieval," from the previous paragraph, in two ways. It can mean how much would be learned relative to doing nothing, which is how we meant it. Or it can mean how much *more* would be learned as a result of retrieval relative to a restudy control condition, which is how the term is usually used. As we illustrate in Fig. 4, this distinction is crucial. The boxes in Fig. 4 contain hypothetical data that we created to illustrate a point. The value in each box represents an amount of learning (ie, an increase in storage

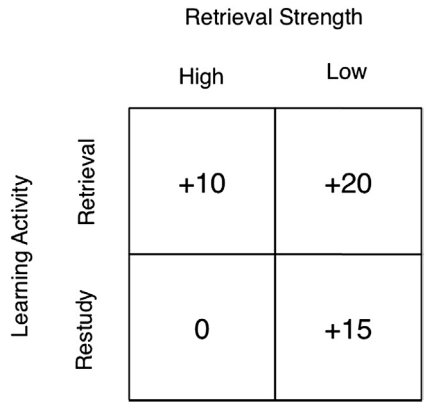

Figure 4 Hypothetical amounts of learning that would accrue a result of four types of trial. As the figure shows, even if retrieval produces more learning for weak items than strong items (top right is greater than top left), the benefit of retrieval over restudy might show the opposite pattern (top minus bottom on the right is less than top minus bottom on the left).

strength) that might accrue as a result of a learning trial. The figure is consistent with NTD in the sense that (1) retrieval results in more learning than restudy (top row versus bottom row) and (2) more learning occurs when retrieval strength is low than when it is high (left column versus right column). However, when we examine the advantage of retrieval over restudy, we reach the opposite conclusion: There is a bigger advantage of retrieval when retrieval strength is high (top minus bottom, left column) than when it is low (top minus bottom, right column). In other words, paradoxically, pretesting could actually lead to more learning than standard testing, and at the same time, data comparing retrieval to presentation could make pretesting appear to be less effective than learning, not due to differences in the retrieval but due to differences in the effectiveness of the control condition.

The numbers in Fig. 4 were constructed to make a point about different ways of measuring the value of retrieval. We have no empirical evidence to support those numbers and we do not claim they match reality. There is some indication in the literature that the pattern of data in Fig. 4 is not unrealistic, however. For example, Karpicke and Roediger (2008) found that after a correct response—when retrieval strength is high (left column)—restudy produced almost no learning but retrieval produced quite a lot. And with respect to low retrieval strength (right column), in this chapter we have shown the advantage of retrieval over presentation, but of course

when retrieval strength is low restudy can produce significant learning. More research is needed to investigate the possibility that Fig. 4 is an accurate depiction of how learning works.

Notice that we are now saying that perhaps retrieval success versus failure *does* predict learning. How can this be if retrieval success does not play a causal role in learning? The causal relationship might work as follows: When an item has relatively low retrieval strength, (1) retrieval failure is more likely and (2) more subsequent learning should occur. Thus, retrieval success versus failure should be correlated with the amount of subsequent learning. This is the classic third variable problem: there is a correlation between learning and retrieval success/failure, but a third variable, retrieval strength, is the cause of the relationship. (There are umpteen such causal relationships; for example, realizing that life moves pretty fast can cause teenagers to aggravate their high school principals more and to take more joyrides in their parents' cars, even if the latter two variables are not causally related.) In other words, we believe more retrieval failure may be correlated with more learning, but we do not believe failure versus success plays a *causal* role in learning.

8.2 Retrieval Effort Hypothesis

In addition to the desirable difficulties framework (eg, Bjork, 1994, 1999) and the new theory of disuse (eg, Bjork & Bjork, 2011), the *retrieval effort hypothesis* (REH; see Pyc & Rawson, 2009) also states that retrieval effort is a moderator of retrieval practice benefits. The primary claim stemming from REH is that successful but effortful retrieval attempts enhance memory more than successful but easy retrieval attempts (consistent with the top row of Fig. 4). Notice that this claim does not directly address the benefits of unsuccessful retrieval practice. Pyc and Rawson (2009) claim that: "The general principles of the [desirable difficulties] framework specify that within any learning task or domain, difficult but successful processing will be better for memory than difficult but unsuccessful processing, a relatively intuitive claim." (p. 437). It is perhaps disputable as to whether this is an intuitive claim made by the desirable difficulties framework, given that Bjork (1999) stresses that learning (ie, what you gain from a learning episode) often occurs without clear improvements in performance (ie, how well you perform on a test). Either way, if success is defined as successfully retrieving a memory, we have to disagree with Pyc and Rawson's quote, but if successful processing includes failing to retrieve and then being told the answer, then their claims seem consistent with ours.

The more important point is that Pyc and Rawson (2009) investigated items that were successfully retrieved during practice with more versus less retrieval effort (as measured via first key press latencies on a correct recall trial; see p. 441), but did not examine effortful but unsuccessful retrieval attempts. REH states that more retrieval effort enhances the benefits of retrieval practice when successful, but it is less clear what REH predicts with respect to more versus less retrieval effort when the retrieval attempt is unsuccessful. We believe that REH could easily be extended to incorporate unsuccessful retrieval attempts and, if so, it would predict the same thing for unsuccessful retrieval attempts as it already predicts for successful ones: that more retrieval effort leads to more learning. In fact, we expect it might make predictions in line with Fig. 4.

8.3 Elaborative Retrieval Hypothesis

We now turn to the first of three theories that propose a mechanism (ie, set of processes) that underlie and explain the benefit of retrieval. The *elaborative retrieval hypothesis* (ERH; see Carpenter, 2009, 2011) states that when one is presented with a cue (eg, frog) and attempts to retrieve a target (eg, pond), the retrieval attempt activates semantic information related to the cue. In the example above, "frog" might activate words like, "green," "water," "lily pad," and "lake." This semantic information is the key mechanism by which testing promotes learning. As a result of the retrieval attempt (and subsequent feedback, if necessary) the direct connection between frog—pond is activated, but so are mediated connections, such as frog—water—pond and frog—lily pad—pond. Critically, these mediated pathways enhance the likelihood of future recall because any of the mediating information can now serve as a potential route to the target.

Carpenter (2011) showed support for this hypothesis by demonstrating that testing improved cue—target learning (eg, *Mother: Child*) relative to studying, and more importantly, it also improved performance when the final test used mediating words, which were never presented initially, such as the cue word (eg, *Father: ????*), which would be a cue to retrieve *Child*. The idea is that the *Father—Child* connection was activated when participants were shown *Mother: ???* and asked to recall *Child*. ERH does not make a prediction about retrieval success versus retrieval failure, but it seems capable of explaining how learning works in both cases. Activation of mediators could happen during the retrieval attempt, and then, when the answer becomes available (whether via retrieval success or external feedback) a process begins of strengthening active connections, including the direct

cue—target connection as well as indirect cue—mediator—target connections. There is no a priori reason why retrieval success versus failure needs to matter in ERH.

8.4 Search Set Theory

Grimaldi and Karpicke (2012) proposed a *search set theory* to explain why pretesting improves subsequent memory. Search set theory is similar to ERH, in that test trials are purported to activate information related to the cue word (eg, tide—???? activates potential candidate solutions such as *beach*, *surf*, and *ocean*). Search set theory assumes that one of the activated traces is the target word itself (eg, if the word pair is tide—beach, *beach* is presumed to be activated along with other potential candidates). Activating the correct target word during a pretest trial (eg, *tide* activates *beach*) helps subsequent encoding of the cue—target pair when feedback is provided (eg, tide—beach). If the target word is not activated during a pretest trial, then pretesting will fail to enhance subsequent encoding. Support for search set theory comes from the fact that most prior studies showing the benefits of pretesting used related materials (in which the chance of activating the correct word on a pretest trial is high), with minimal benefits showing up when the word pairs were unrelated (in which the chance of activating the correct word on a pretest trial is low; eg, Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012).

With respect to retrieval success, search set theory would seem to claim that as long as the trace is activated during the pretest trial, pretesting improves memory. For instance, a participant may activate many traces related to *tide* (eg, *beach*, *wave*, *surf*, and *ocean*); however, they may incorrectly answer with *beach* when the correct answer is *wave*. But this might not matter, as trace activation is all that is necessary to benefit from a failed test trial (and not necessarily typing in the correct answer). However, search set theory is difficult to test because it hinges on which traces are activated during a pretest trial, but these trace activations are not typically measured in research on retrieval. The typical procedure is to give participants feedback following an unsuccessful retrieval attempt, not to stop, and measure priming in their semantic network. Thus, further research is needed to evaluate this theory.

The main differences between search-set theory and ERH are (1) search set theory claims that the trace is activated during an unsuccessful memory search and (2) in search set theory, retrieval effects are attributed primarily to the strength of the direct association between the cue and target, whereas ERH hinges on mediated connections between the cue and target. Like

ERH, however, search-set theory seems compatible with the idea that successful and unsuccessful retrieval attempts can be explained by a single mechanism.

8.5 Episodic Context Account

Lehman, Smith, and Karpicke (2014) describe an *episodic context account* (ECA) to explain the benefits of testing. Contextual features are processed and stored within a memory, and these contextual features change over time. During retrieval, learners try to recreate or reinstate a prior learning context, and each retrieval attempt causes the contextual representation to be updated. For example, studying an item in Context A, and retrieving that item in Context B results in a composite representation of both contexts. Studying in Context A and then *restudying* in Context B, on the other hand, is less effective because restudying does not elicit a memory of Context A, and thus no composite trace combining the contexts is formed. Repeatedly retrieving information results in an even more diverse and varied contextual representation. The reason retrieval enhances memory is that a richer contextual trace is more likely to overlap with the contexts present when one needs the memory (eg, during the final test) and the context cues that are present help elicit the context cues associated with the target, making the target easier to retrieve. Thus, according to ECA, the mechanism underlying the benefit of retrieval is enhanced contextual representations.

ECA makes some admirably specific predictions. For example, retrieval should be more beneficial when contextual representations present at the time of retrieval practice are more heterogeneous, because a variety of encoding contexts create a richer composite episodic trace. Because presentation trials integrate contextual information less effectively, the advantage of retrieval over presentation should be larger when retrieval contexts are more varied. Testing this hypothesis would be one way to test of the viability of ECA.

The other theories we have discussed are compatible with our proposal that retrieval success and retrieval failure enhance learning because of the same underlying mechanism. The same may not be true of ECA. ECA easily explains the learning that occurs when retrieval is successful. The same is true in a potentiation procedure (Fig. 2, left panel) because there are multiple retrieval attempts; for example, being asked a question for the second time might activate (consciously or unconsciously) the context when/where one heard the question for the first time, and so forth. But a retrieval attempt

cannot reinstate a prior context if there is no prior context, so it is unclear how ECA would explain the benefits of the pretesting procedure (Fig. 2, right panel). Of course, even if ECA plays no role in the benefit of pretesting, it might still play an important role in the benefit of retrieval in other situations (ie, when the information has been studied previously). Notice, also, that under the conditions necessary for ECA to apply, its mechanism does not necessitate making a distinction between learning from retrieval success and learning from retrieval failure.

9. CONCLUSION

Our conclusions can be summarized in one sentence: Retrieval attempts result in either success or failure, but both enhance learning and there is no evidence that they produce different amounts of learning or work via different mechanisms. A small boom in research on unsuccessful retrieval attempts in the last five years has created a strong evidence base showing benefits across a wide variety of materials (eg, trivia questions, cue—target pairs, math problems, face—age associations, etc.), for a wide variety of learners (eg, both younger and older adults, and even some older adults with memory impairments), and across a variety of delays (eg, both on immediate and delayed final tests). Furthermore, these benefits have been extended from the laboratory to the classroom (eg, Kapur & Bielaczyc, 2012). Evidence that retrieval success and retrieval failure are equally effective is not as abundant, but a small amount has accumulated in the form of a recent study by Kornell et al. (2015) as well as data reported in this chapter.

9.1 Theoretical Implications

With respect to the mechanism underlying retrieval, we believe trying to think of an answer potentiates learning that occurs when the answer becomes available, and this basic process can apply in almost exactly the same way regardless of whether or not retrieval is successful. How the answer comes to mind—whether via retrieval success or external feedback—may be a surface feature that does not actually impact the process of learning. However, the idea that there are two stages of processing says nothing about the key question: What processes occur during these two stages to make retrieval beneficial? We outlined a number of theories that set forth plausible processes. There is no reason why these theories should be mutually exclusive. Indeed, based on a mix of intuition and data collected

in our lab, we believe ERH is right about mediators playing an important role and search-set theory is right that retrieval probably enhances direct cue—target connections as well. The degree to which context plays a role in the benefit of retrieval is an empirical question (ie, more support is needed for ECA). More evidence is needed to test all of these theories.

There are other unanswered questions as well. One intriguing hypothesis is that feedback following a retrieval attempt might have an inhibitory effect on incorrect answers (Carrier & Pashler, 1992; Kornell et al., 2009). For example, imagine a retrieval attempt based on the cue *Club*. During Stage 1 related concepts like country, golf, or breakfast might come to mind, but if the target was revealed to be *Caveman*, these concepts, which are related to the wrong meaning of the word club, might be inhibited. Inhibition effects tend to be strongest for items that compete with the actual answer (eg, *Fruit—???* followed by *Apple* might inhibit *Pear* but it probably would not inhibit *Vegetable*) so a limited subset of responses might be inhibited (Anderson, Bjork, & Bjork, 1994).

Another intriguing question has to do with activation of related information. Search-set theory and ERH both predict that concepts related to the cue will be activated more during retrieval than during presentation. Carpenter (2011) tested this hypothesis by presenting a mediator and asking participants to come up with targets, but we would like to see research in which the mediator is not presented and the question is whether the related mediators themselves have become primed by act of retrieval.

We also believe future research should endeavor to test the role of mediators directly. Carpenter (2011) showed that following retrieval, participants who were shown a mediator were more likely to respond with the target. But in research on retrieval, the standard final test procedure involves presenting the cue and asking for the target. We believe that an experiment that used the standard paradigm, and showed that mediators are integral to the value of retrieval, would represent a significant increase in the evidence for ERH.

9.2 Practical Implications

It is important to raise a practical point: Our review has addressed the effects of retrieval success versus failure when they are isolated from other factors (such as item difficulty). Understanding the specific, unconfounded effects of retrieval success is important for theories of retrieval. But it is not necessarily important for real life. It is exceedingly rare in education or everyday life for retrieval success to be manipulated by itself. For example, if a teacher

were told that retrieval success does not matter, he would not be able to manipulate retrieval success holding other factors constant. If he decided to make changes that would decrease retrieval success, he would do so by asking his students harder questions. In other words, in real life, what would change would be retrieval difficulty. As we have said, retrieval difficulty can have large effects on learning even if retrieval success per se does not (see Fig. 4). Thus, whether or not retrieval success per se affects learning is an academic question, not one that pertains to real life. In short, the answer to the second question we asked in this chapter is that retrieval success versus failure, per se, does not seem to affect learning, but we do not recommend making decisions about how to teach, or learn, based on this finding.

The point we hope teachers and learners will take away from our review is the answer to our first question: Retrieval attempts *do* enhance learning even when they are not successful. Students and teachers are prone to actively seek strategies that safeguard retrieval success, or to avoid strategies that might stimulate retrieval failure. We believe these efforts are often misguided. Difficult retrieval is a very effective way to study. The danger that consistent retrieval failure will undermine students' motivation is real and should not be taken lightly; ideally, though, students can learn to accept struggle as part of learning. Instead of worrying about retrieval success, students and teachers should embrace errors as a path to knowledge.

REFERENCES

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063–1087.

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.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ: Lawrence Erlbaum Associates.

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, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Routledge.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, 56–64.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin and Review, 18*(6), 1238–1244. http://dx.doi.org/10.3758/s13423-011-0173-y.

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

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*(5), 826–830.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition, 20*(6), 633–642.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380.

Clare, L., & Jones, R. S. (2008). Errorless learning in the rehabilitation of memory impairment: a critical review. *Neuropsychology Review, 18*(1), 1–23.

Cyr, A. A., & Anderson, N. D. (2014). Mistakes as stepping stones: effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 841–850.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology, 6*(3), 217–226.

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory and Cognition, 38*(7), 951–961.

Glisky, E. L., & Schacter, D. L. (1989). Extending the limits of complex learning in organic amnesia: computer training in a vocational domain. *Neuropsychologia, 27*(1), 107–120.

Glisky, E. L., Schacter, D. L., & Tulving, E. (1986). Learning and retention of computer-related vocabulary in memory-impaired patients: method of vanishing cues. *Journal of Clinical and Experimental Neuropsychology, 8*(3), 292–312.

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

Guthrie, E. (1952). *The psychology of learning* (Rev. ed.). New York: Harper.

Haslam, C., Gilroy, D., Black, S., & Beesley, T. (2006). How successful is errorless learning in supporting memory for high and low-level knowledge in dementia? *Neuropsychological Rehabilitation, 16*(5), 505–536. http://dx.doi.org/10.1080/09602010500231867.

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 290–296.

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory and Cognition, 40*(4), 514–527.

Izawa, C. (1969). Comparison of reinforcement and test trials in paired-associate learning. *Journal of Experimental Psychology, 81*(3), 600–603.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*(2, Pt. 1), 340–344.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory and Cognition, 30*(6), 823–840.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21*(1), 45–83.

Karpicke, J. D., & Roediger, H. L., III (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968. http://dx.doi.org/10.1126/science.1152408.

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.

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297–1317.

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., 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.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1787–1794. http://dx.doi.org/10.1037/xlm0000012.

McGillivray, S., & Castel, A. D. (n.d.). Memory for age–face associations in younger and older adults: The role of generation and schematic support.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3–8. http://dx.doi.org/10.1037/0278-7393.31.1.3.

Pashler, H., Zarow, G., & Triplett, B. (2003). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1051–1057.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*(2), 644–667.

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.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: how much is enough? *Journal of Experimental Psychology: General, 140*(3), 283–302.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*(3), 243–257.

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., III, & Karpicke, J. D. (2006a). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x.

Roediger, H. L., & Karpicke, J. D. (2006b). 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.

Skinner, B. F. (1968). *The technology of teaching*. Englewood Cliffs, NJ:: Prentice-Hall.

Squires, E. J., Hunkin, N. M., & Parkin, A. J. (1997). Errorless learning of novel associations in amnesia. *Neuropsychologia, 35*(8), 1103–1111.

Terrace, H. S. (1963). Errorless transfers of a discrimination across two continua. *Journal of the Experimental Analysis of Behavior, 6*, 223–232.

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory what aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*(9), 1127–1131.
Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin and Review, 19*(5), 899–905.
Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: evidence for associative asymmetry. *Journal of Memory and Language, 75*, 14–26.
Wilson, B. A., Baddeley, A., Evans, J., & Shiel, A. (1994). Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological Rehabilitation, 4*(3), 307–326.